

Introduction to Data Science

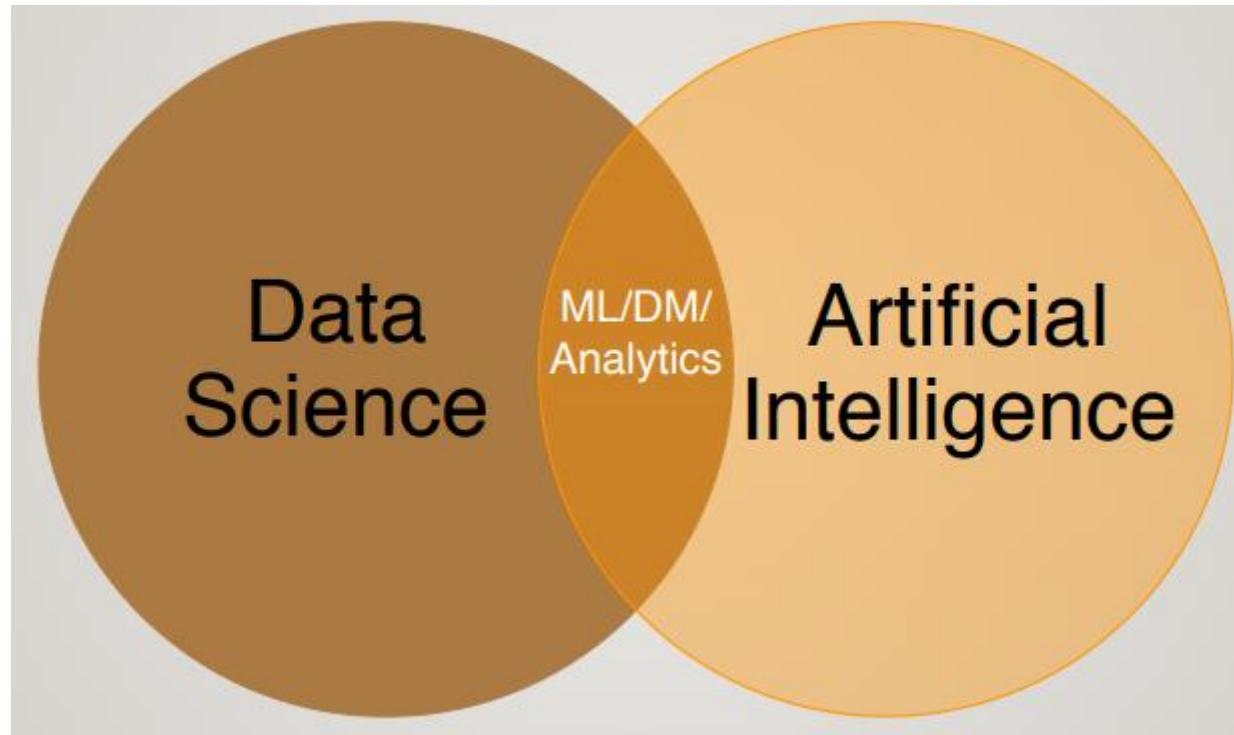
Define

- “Data science, also known as data-driven science, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured.”
- It combines aspects of statistics, machine learning, and domain expertise to analyze data and make informed decisions.

DATA SCIENCE AND BIG DATA

- They are not the “same thing”
- **Big data = crude oil**
- Big data is about extracting “crude oil”, transporting it in “mega tankers”, siphoning it through “pipelines”, and storing it in “massive silos”
- Data science is about refining the “crude oil”

DATA SCIENCE AND ARTIFICIAL INTELLIGENCE



Key Concepts

- **Data:** The raw material of data science. It can be structured (organized in tables or databases), semi-structured (like emails or social media posts), or unstructured (images, videos, audio).
- **Analysis:** The process of examining data to identify patterns, trends, and relationships. This can involve statistical methods, machine learning algorithms, and data visualization techniques.
- **Insights:** The valuable information extracted from data analysis. Insights can be used to make better decisions, improve products and services, and gain a competitive advantage.

Key Concepts

The field of data science typically involves three key areas:

- **Data collection and processing:** This process, also referred to as data preparation, involves gathering data from various sources and cleaning it to ensure accuracy and reliability. The collected data may come from databases, spreadsheets, online sources, and other types of data storage systems.
- **Data analysis:** Data scientists use statistical methods and machine learning algorithms to explore and analyze the data. This step helps to identify patterns, correlations, trends, and other insights hidden in the data.
- **Data interpretation and communication:** Once the analysis is complete, data scientists interpret the results and communicate them to stakeholders in a way that's easy to understand. This often involves creating visualizations, reports, and presentations.

Importance of data science

In Various Industries

From healthcare to finance, data science is playing a crucial role in various industries. By analyzing large amounts of data, businesses can reduce costs, increase efficiency, and improve customer satisfaction. In this section, we will explore the benefits of data science in different sectors.



Benefits of data science for businesses

Better Decision Making

Data science helps businesses make better decisions based on insights gained from analyzing large amounts of data.

Increased Efficiency

Data science allows businesses to automate processes, reduce costs, improve efficiency, and streamline operations.

Improved Customer Experience

By analyzing customer data, businesses can personalize their offerings, improve customer satisfaction, and drive customer retention.



Applications of Data Science

- **Business:** Customer segmentation, fraud detection, personalized recommendations, market research, risk assessment.
- **Healthcare:** Disease diagnosis, drug discovery, personalized medicine, medical imaging analysis, patient monitoring.
- **Finance:** Algorithmic trading, risk management, credit scoring, fraud detection, financial forecasting.
- **Social Media:** Sentiment analysis, recommendation systems, targeted advertising, network analysis.
- **Government:** Predictive policing, disaster response, urban planning, public health surveillance.

Data Science Challenges

1. Data Quality Issues:

- **Inaccurate Data:** Errors, inconsistencies, and outdated information can lead to flawed analyses and misleading conclusions.
- **Incomplete Data:** Missing values can hinder analysis and reduce the effectiveness of models.

2. Data Volume and Velocity:

- **Big Data:** The sheer volume of data generated today can be overwhelming to process and analyze efficiently.
- **Real-time Data:** The need to analyze streaming data in real-time presents challenges for data processing and model deployment.

Data Science Challenges

3. Data Privacy and Security:

- **Data Breaches:** Sensitive data is vulnerable to cyberattacks, leading to potential harm to individuals and organizations.
- **Regulations:** Compliance with data privacy regulations (e.g., GDPR, CCPA) adds complexity and constraints to data usage.

4. Lack of Skilled Professionals:

- **Talent Gap:** There is a significant shortage of skilled data scientists with the necessary expertise in areas like machine learning, statistics, and programming.
- **Interdisciplinary Skills:** Finding professionals with both technical skills and domain expertise is challenging.

Data Science Challenges

5. Explainability and Interpretability:

- **Black Box Models:** Many advanced machine learning models are complex and difficult to understand, making it hard to explain their decisions and build trust.
- **Bias and Fairness:** Models can inadvertently reflect biases present in the training data, leading to unfair or discriminatory outcomes.

6. Communication and Collaboration:

- **Bridging the Gap:** Effectively communicating complex technical concepts to non-technical stakeholders is crucial but often challenging.

7. Ethical Considerations:

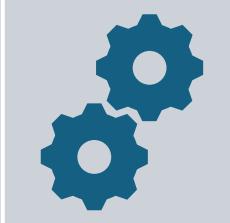
- **Transparency and Accountability:** Ensuring transparency and accountability in the use of data and the development of AI systems.

Introduction to Data Science

Software Engineering for Data Science



Bridges the gap between data science and software development.



It involves **applying software engineering principles & practices to the data science lifecycle**, ensuring that data-driven solutions are not only effective but also efficient, scalable, and maintainable.

Software Engineering for Data Science

Key Concepts

- **DataOps:** A set of practices that aim to shorten the system development lifecycle while delivering features, fixes, and updates frequently and reliably.
 - It focuses on the entire data pipeline, from data ingestion and transformation to analysis and visualization.
- **MLOps:** A set of practices that automate and streamline the machine learning (ML) lifecycle, enabling faster and more reliable development and deployment of ML models.
 - It covers the entire ML workflow, from data preparation and model training to deployment, monitoring, and retraining.

Software Engineering for Data Science

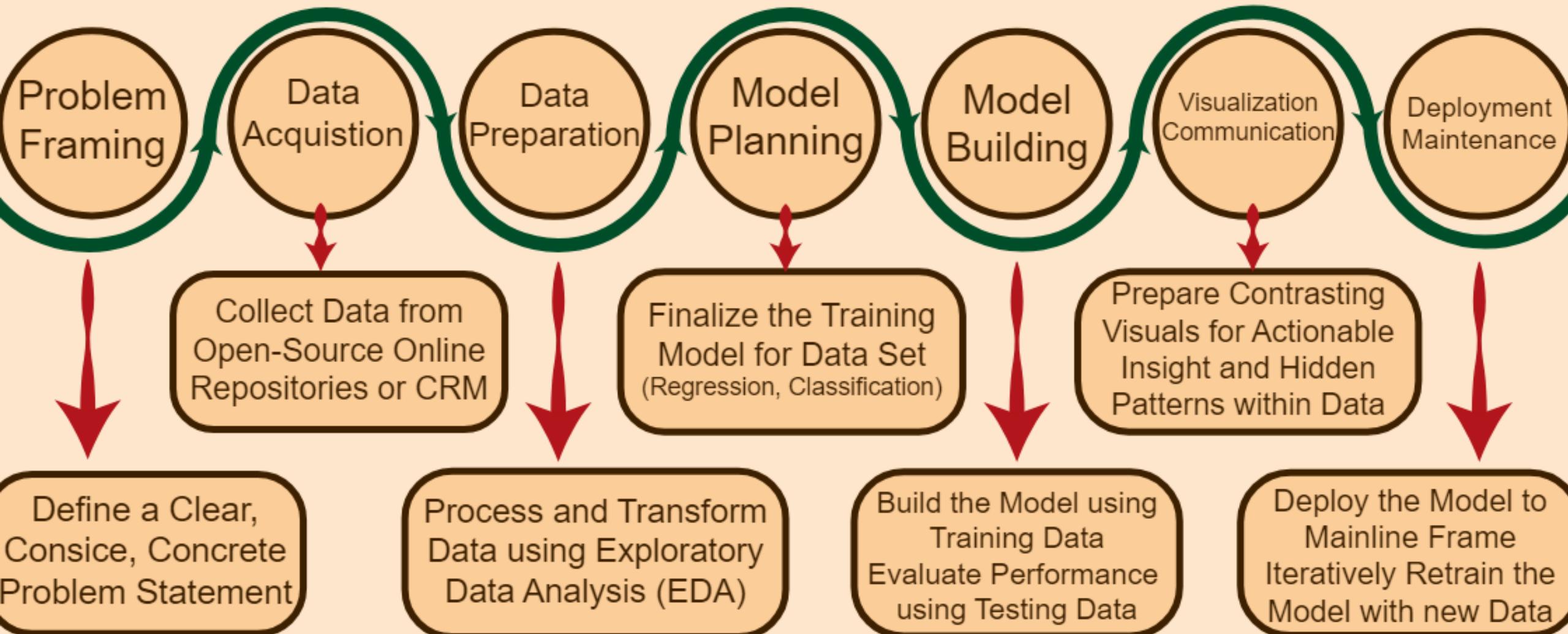
Core Principles

- **Version Control:** Using tools like Git to track changes in code, data, and models.
- **Testing and Quality Assurance:** Implementing unit tests, integration tests, and end-to-end tests to ensure the accuracy and reliability of data pipelines and ML models.
- **Continuous Integration and Continuous Delivery (CI/CD):** Automating the build, test, and deployment processes to accelerate the development cycle.
- **Infrastructure as Code:** Defining and managing infrastructure (e.g., servers, databases, cloud resources) using code, making it easier to reproduce and scale.
- **Scalability & Performance:** Designing systems that can handle large volumes of data and high-throughput workloads.
- **Reproducibility:** Ensuring that experiments and models can be easily reproduced to validate results and maintain consistency.
- **Collaboration:** Fostering collaboration between data scientists, software engineers, and other stakeholders.

Software Engineering for Data Science

Aspect	MLOps	DataOps
Automation	Automates ML model deployment and monitoring.	Automates data pipeline processes.
Collaboration	Encourages teamwork between data scientists and engineers.	Emphasizes collaboration across data teams to achieve common goals.
CI/CD	Uses CI/CD to deploy ML pipelines and update ML models.	Implements CI/CD practices for data pipeline deployment
Model Cataloging	Catalogs ML model versions and associated artifacts.	Catalogs data versions and metadata.
Version Control	Tracks code and model versions for consistency and review.	Tracks data versions for auditability.
Monitoring	Monitors ML models for performance and bugs.	Monitors data pipelines for issues and errors.
Governance	Ensures compliance with regulations like GDPR and HIPAA.	Ensures data quality and compliance with regulations.
DevOps Principles	Draws inspiration from DevOps for automation and teamwork.	Draws inspiration from DevOps for collaboration and innovation.

DATA SCIENCE PROCESS



Data Science Process Roles



Data Scientist: The core role responsible for analyzing data, building and evaluating models, and extracting meaningful insights. They possess strong statistical, mathematical, and programming skills.



Data Engineer: Focuses on building and maintaining the data infrastructure, including data pipelines, data warehouses, and data lakes. They ensure data quality, availability, and accessibility for analysis.



Data Analyst: Gathers, cleans, and prepares data for analysis. They perform exploratory data analysis (EDA) and generate reports and visualizations to communicate insights to stakeholders.

Data Science Process Roles



Machine Learning Engineer: They ensure model performance, scalability, and reliability.



Business Analyst: Understands business needs and translates them into data science problems. They act as a bridge between the business and the data science team.



Data Architect: Designs and implements data solutions, including data models, databases, and data integration strategies

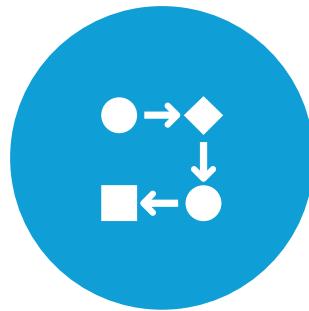
Types of Data Analytics



DESCRIPTIVE (BUSINESS
INTELLIGENCE AND DATA
MINING)



PREDICTIVE (FORECASTING)



PRESCRIPTIVE
(OPTIMIZATION AND
SIMULATION)



DIAGNOSTIC ANALYTICS

Descriptive Analytics

- It looks at data and analyses past event for insight as to how to approach future events.
- Descriptive analytics looks at past performance and understands the performance by mining historical data to understand the cause of success or failure in the past.
- Almost all management reporting such as sales, marketing, operations, and finance uses this type of analysis.
- Examples of Descriptive analytics are company reports that provide historic reviews like: Data Queries, Reports, Descriptive Statistics, Data dashboard

Predictive Analytics

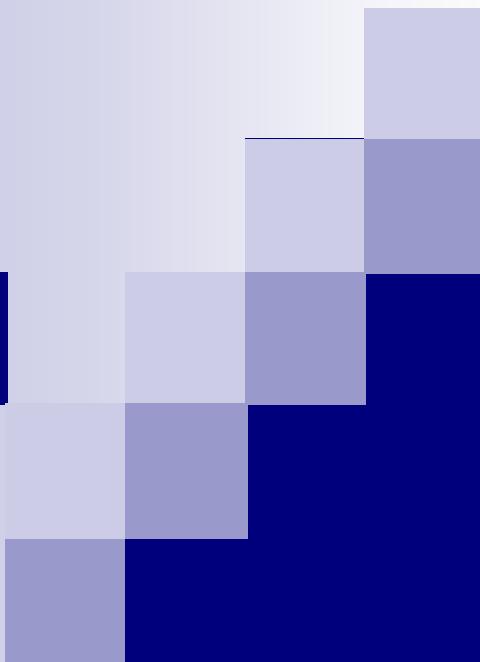
- Predictive analytics turn the data into valuable, actionable information.
- It uses data to determine the probable outcome of an event or a likelihood of a situation occurring.
- Predictive analytics holds a variety of statistical techniques from modeling, machine, learning, data mining, and game theory that analyze current and historical facts to make predictions about a future event.
- Techniques that are used for predictive analytics are:
 - Linear Regression
 - Time series analysis and forecasting
 - Data Mining

Prescriptive Analytics

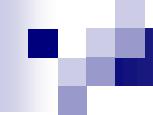
- Prescriptive analytics goes beyond predicting future outcomes by also suggesting action benefits from the predictions and showing the decision maker the implication of each decision option.
- For example, Prescriptive Analytics can benefit healthcare strategic planning by using analytics to leverage operational and usage data combined with data of external factors such as economic data, population demography, etc.
- This type of analytics talks about an analysis that is based on rules and recommendations, to prescribe a certain analytical path for an enterprise.
- At the next level, prescriptive analytics will automate decisions and actions—how can we make that happen?

Diagnostic Analytics

- In diagnostic analytics, most enterprises start to apply data analytics to answer diagnostic questions such as how and why something happened.
- Some may also call this behavioural analytics.
- Diagnostic analytics is about looking into the past and determining why a certain thing happened. This type of analytics usually revolves around working on a dashboard.
- Use historical data over other data to answer any question or for the solution of any problem. We try to find any dependency and pattern in the historical data of a particular problem.



Introduction to Data Science



Data Science Process

Different methodologies used in data science

- CRISP-DM Methodology,
- SEMMA,
- BIG DATA LIFE CYCLE,
- SMAM.

CRISP-DM: Methodology

CRISP-DM stands for Cross-Industry Standard Process for Data Mining.

- Widely adopted methodology
- Provides a structured approach for planning & executing DM projects.
- Designed to be adaptable across various industries and applications.
- Key Characteristics of CRISP-DM
 - **Iterative:** The process is not strictly linear. You may need to revisit previous phases as you progress.
 - **Flexible:** It can be adapted to various project sizes and complexities.
 - **Industry-Neutral:** Applicable across different domains and sectors.
 - **Focus on Business Value:** Emphasizes understanding business needs and aligning data mining efforts accordingly.

CRISP-DM: Data Mining Operations

1. Business Understanding:

1. Determine business objectives and requirements.
2. Assess situation and resources.
3. Determine data mining goals.

2. Data Understanding:

1. Collect initial data.
2. Describe data.
3. Explore data.
4. Verify data quality.

3. Data Preparation:

1. Select and Clean data.
2. Construct data.
3. Integrate data.
4. Format data.

4. Data Modeling:

1. Select modeling techniques.
2. Generate test design.
3. Build and Assess models.

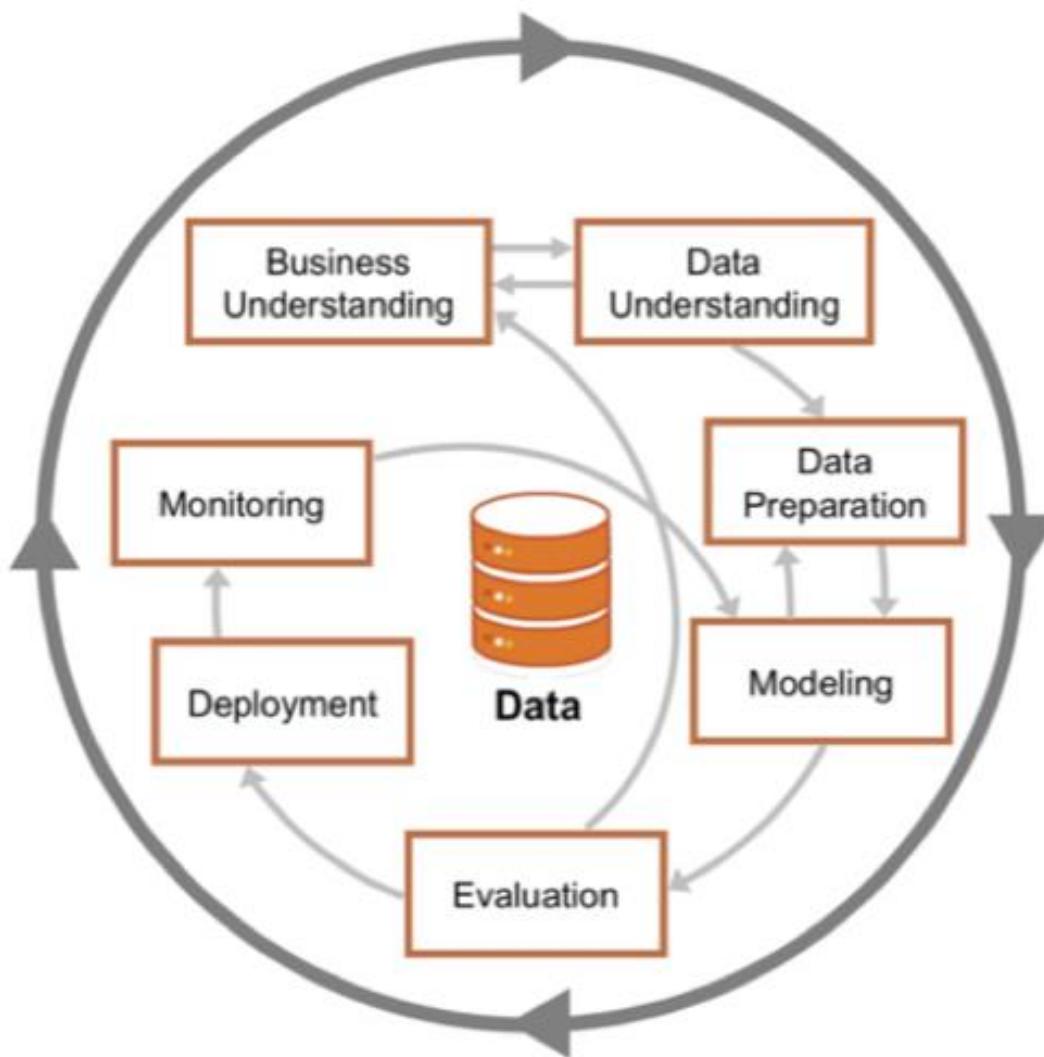
5. Evaluation:

1. Evaluate results.
2. Review process.
3. Determine next steps.

6. Deployment:

1. Plan deployment.
2. Plan monitoring and maintenance.
3. Produce final report.
4. Review project.

CRISP-DM: Methodology



SEMMA

SEMMA is a data mining methodology developed by the SAS Institute. It outlines a five-step process for extracting meaningful insights from data:

1. Sample:

- **Select a representative subset of the data** for analysis. Necessary to manage the computational complexity of working with large datasets.
- Sampling techniques can include random sampling, stratified sampling, and cluster sampling.

2. Explore:

- **Conduct exploratory data analysis (EDA)** to understand the characteristics of the data.
- This involves visualizing the data, identifying patterns, and detecting anomalies.
- Common EDA techniques include histograms, scatter plots, box plots...

SEMMA

3. Modify:

- **Transform and prepare the data for modeling.**
 - **Data cleaning:** Handling missing values, outliers, and inconsistencies.
 - **Feature engineering:** Creating new variables or transforming existing ones.
 - **Data transformation:** Scaling or normalizing data to improve model accuracy.

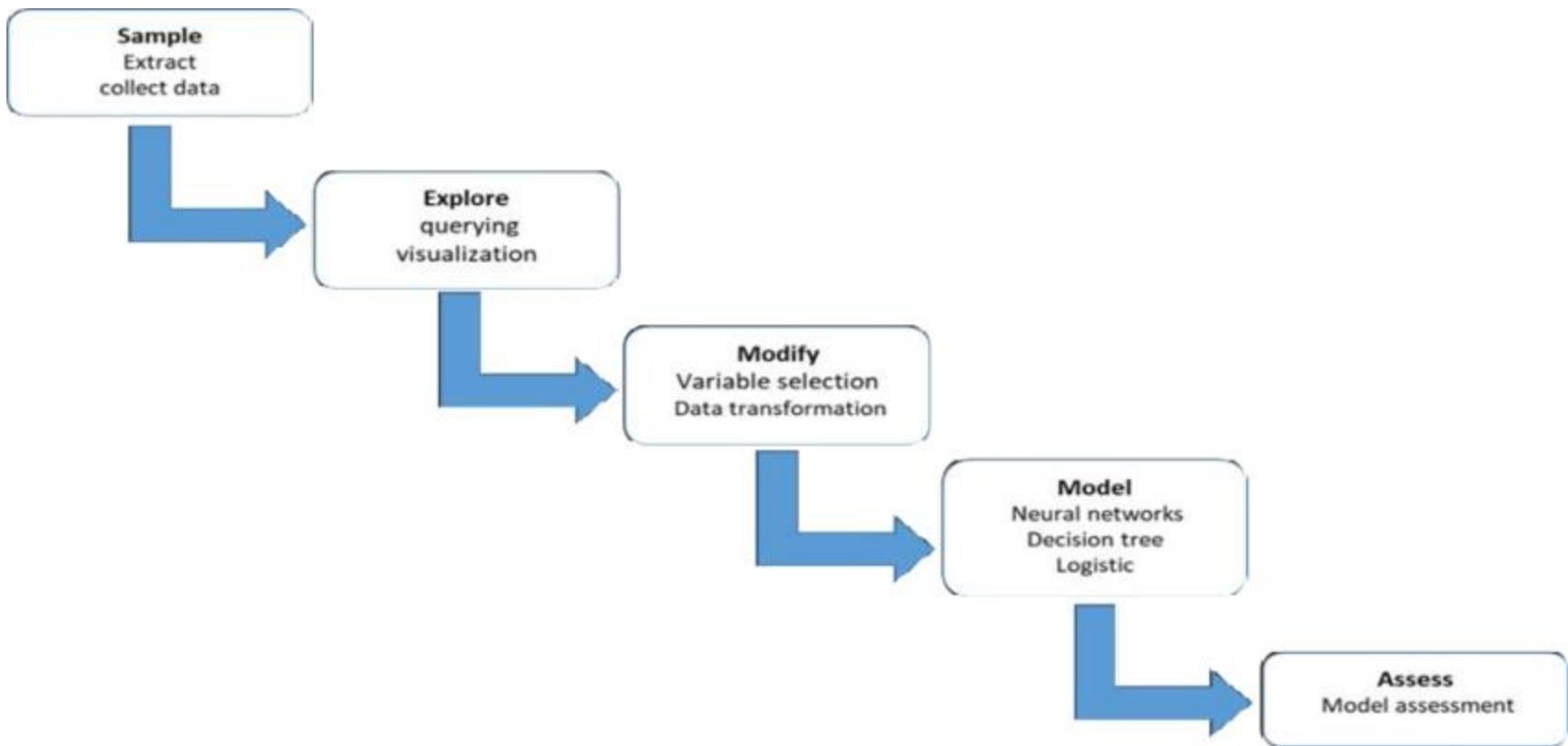
4. Model:

- **Build and train predictive models** using appropriate algorithms.
 - **Regression:** Predicting continuous values.
 - **Classification:** Predicting categorical values.
 - **Clustering:** Grouping similar data points together.

5. Assess:

- **Evaluate the performance of the models** using appropriate metrics.
- This helps to determine the accuracy, reliability, and generalizability of the models.
- Common evaluation metrics include accuracy, precision, recall, and F1-score.

SEMMA



SMAM

- **SMAM** stands for **Sample, Mine, Assess, Maintain**. It's a simplified data science methodology, particularly useful for initial data exploration and analysis. Here's a breakdown:

1. Sample:

- Select a representative subset of the data.

2. Mine:

- Apply data mining techniques to discover patterns and relationships within the data.

3. Assess:

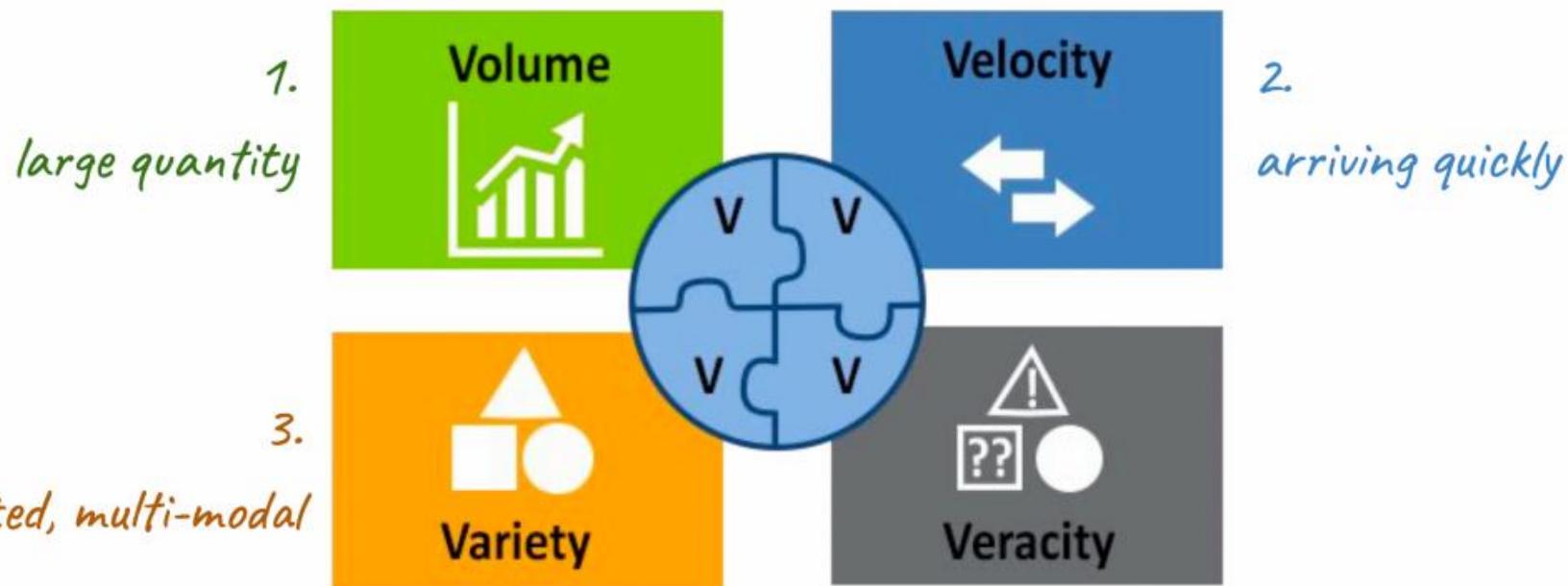
- Evaluate the findings and their implications.

4. Maintain:

- **Update and refine the analysis as new data becomes available.**
- This may involve:
 - **Retraining models** with new data to improve their performance.
 - **Updating data sources** and re-running the analysis.
 - **Incorporating new insights** and adjusting the analysis accordingly.

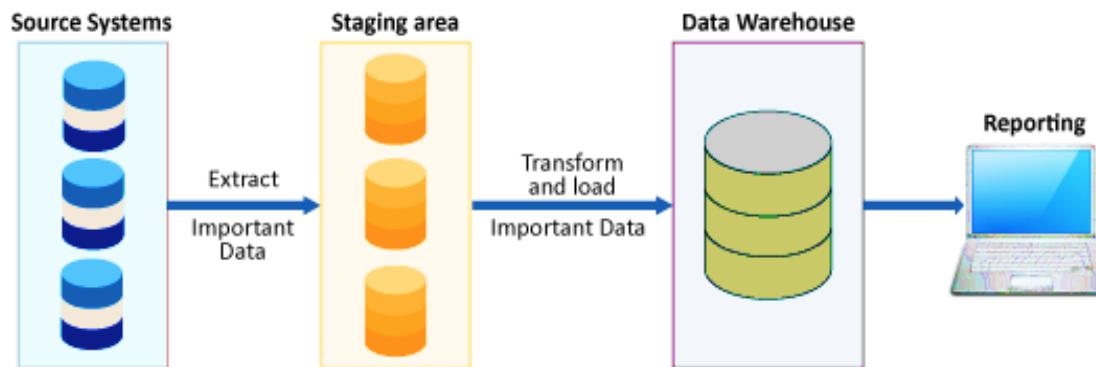
Big Data

ANALYSES WHICH CAN HANDLE THE 3 VS
AND DO IT WITH QUALITY (VERACITY)



ETL vs ELT

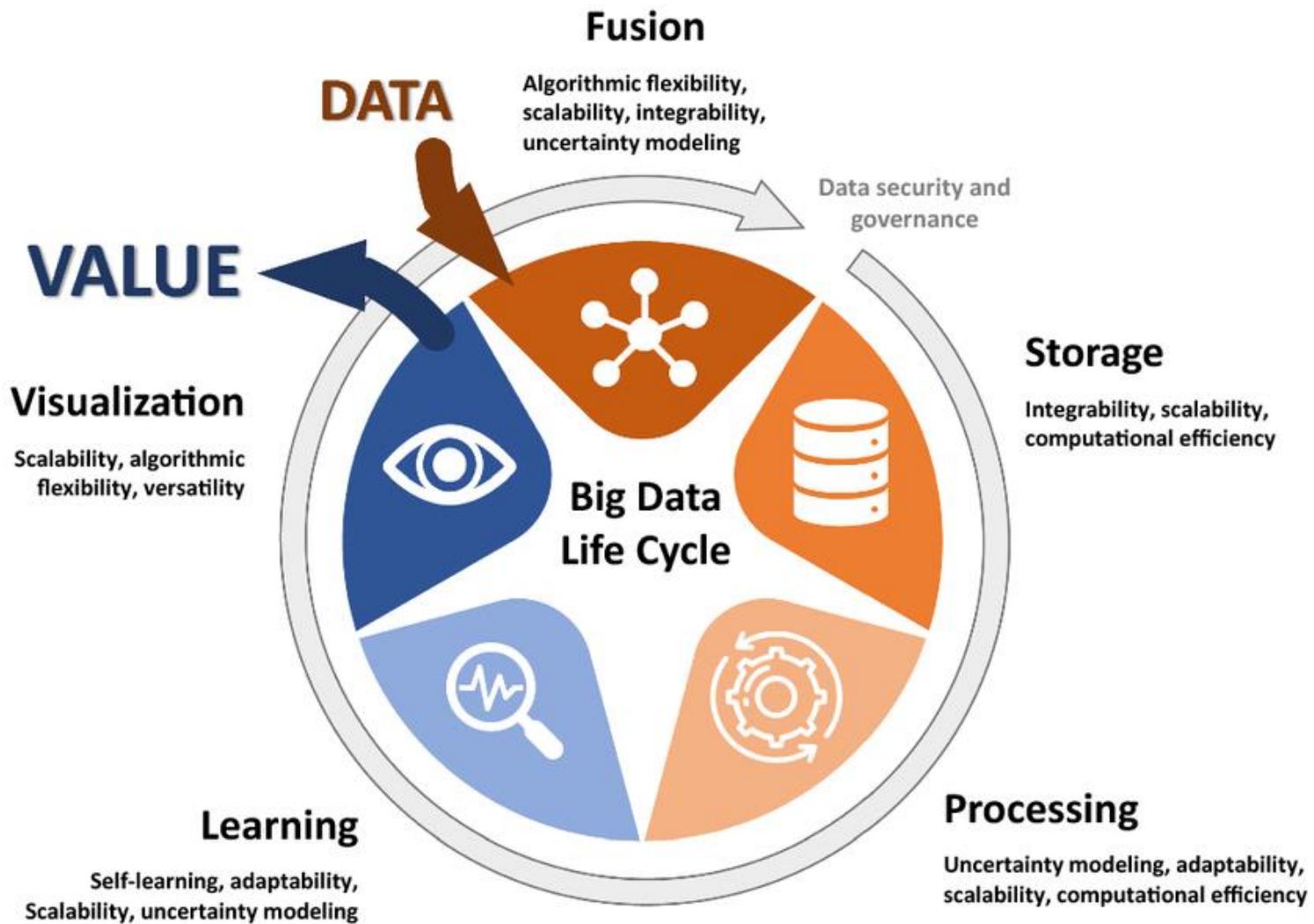
ETL



ELT



Big Data Life Cycle



Statistics

Expectation and Variance of RVs

Discrete Random Variables:

- **Probability Mass Function (PMF):** Let X be a discrete random variable with possible values x_1, x_2, x_3, \dots and corresponding probabilities $p_1 = P(X = x_1), p_2 = P(X = x_2), p_3 = P(X = x_3), \dots$ (where $\sum p_i = 1$). ▼
- **Expectation (Mean):**
$$E[X] = \mu = \sum [x_i * p_i]$$
 (sum over all possible values of x) ▼
- **Variance:**
$$\text{Var}(X) = E[(X - \mu)^2] = \sum [(x_i - \mu)^2 * p_i]$$

$$\text{Var}(X) = E[X^2] - (E[X])^2$$
 (a computationally useful form) ▼
- **Expectation of a Function of X :**
$$E[g(X)] = \sum [g(x_i) * p_i]$$

Expectation and Variance of RVs

General Properties of Expectation:

- **Linearity:** $E[aX + bY] = aE[X] + bE[Y]$, where a and b are constants and X and Y are random variables.
- **Constant:** $E[c] = c$, where c is a constant.

General Properties of Variance:

- **Constant:** $\text{Var}(c) = 0$, where c is a constant.
- **Scaling:** $\text{Var}(aX) = a^2\text{Var}(X)$, where a is a constant.
- **Linear Transformation:** $\text{Var}(aX + b) = a^2\text{Var}(X)$, where a and b are constants. ▼
- **Independence:** If X and Y are independent random variables, then $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. (This does *not* generally hold if X and Y are dependent).

Expectation and Variance of RVs

Q1. Let Z be a random variable with the following probability distribution:

$$P(Z = -1) = 0.2 \quad P(Z = 0) = 0.5 \quad P(Z = 1) = 0.3$$

Define a new random variable $W = Z^2$.

- Find the expected value of W , $E[W]$.
- Find the variance of W , $\text{Var}(W)$.

First, we need to find the probability distribution of W .

- If $Z = -1$, then $W = (-1)^2 = 1$. $P(W = 1) = P(Z = -1) = 0.2$
- If $Z = 0$, then $W = (0)^2 = 0$. $P(W = 0) = P(Z = 0) = 0.5$
- If $Z = 1$, then $W = (1)^2 = 1$. $P(W = 1) = P(Z = 1) = 0.3$

Notice that W can only take the values 0 and 1. The probability distribution of W is:

- $P(W = 0) = 0.5$
- $P(W = 1) = 0.2 + 0.3 = 0.5$

Now we can calculate $E[W]$:

$$\begin{aligned} E[W] &= \sum [w * P(W = w)] \\ &= (0 * 0.5) + (1 * 0.5) \\ &= 0 + 0.5 = 0.5 \end{aligned}$$

$$\text{Var}(W) = E[W^2] - (E[W])^2$$

Since W can only be 0 or 1, W^2 will also only be 0 or 1. In fact, $W^2 = W$ in this case. This is because $0^2=0$ and $1^2=1$. So, $E[W^2] = E[W] = 0.5$

$$\begin{aligned} \text{Var}(W) &= E[W^2] - (E[W])^2 \\ &= 0.5 - (0.5)^2 \\ &= 0.5 - 0.25 \\ &= 0.25 \end{aligned}$$

Expectation and Variance of RVs

Q2. Let X be a random variable with $E[X] = 5$ and $\text{Var}(X) = 2$. Let $Y = 3X - 4$.

- Find $E[Y]$.
- Find $\text{Var}(Y)$.

We can use the linearity of expectation, which states that $E[aX + b] = aE[X] + b$, where 'a' and 'b' are constants.

$$\begin{aligned} E[Y] &= E[3X - 4] \\ &= 3E[X] - 4 \quad (\text{using linearity of expectation}) \\ &= 3(5) - 4 \quad (\text{substituting } E[X] = 5) \\ &= 15 - 4 \\ &= 11 \end{aligned}$$

We can use the property of variance that states $\text{Var}(aX + b) = a^2\text{Var}(X)$, where 'a' and 'b' are constants. Notice that the constant term 'b' does not affect the variance.

$$\begin{aligned} \text{Var}(Y) &= \text{Var}(3X - 4) \\ &= 3^2\text{Var}(X) \quad (\text{using the property of variance}) \\ &= 9(2) \quad (\text{substituting } \text{Var}(X) = 2) \\ &= 18 \end{aligned}$$

Co-Variance of RVs

Q3. Let U and V be two independent standard normal random variables, i.e., $U, V \sim N(0, 1)$. Define the new random variables:

$$R = 5 + 2U - 3UV$$

$$S = 2 - U + V$$

Find $\text{cov}(R, S)$

The covariance between two random variables R and S is defined as:

$$\text{cov}(R, S) = E[(R - E[R])(S - E[S])] = E[RS] - E[R]E[S]$$

Co-Variance of RVs

Covariance:

- Definition: $\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
- Relationship to Variance: $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$ (This holds in general, whether or not X and Y are independent.)
- Independence: If X and Y are independent, $\text{Cov}(X, Y) = 0$. (The converse is not necessarily true.) 

Standard Deviation:

- The standard deviation of X, denoted σ or $\text{SD}(X)$, is the square root of the variance: $\sigma = \sqrt{\text{Var}(X)}$. It provides a measure of the spread of the distribution in the original units of the random variable. 

Co-Variance of RVs

Q1. Let U and V be two independent standard normal random variables, i.e., $U, V \sim N(0, 1)$. Define the new random variables: $R = 5 + 2U - 3UV$ and $S = 2 - U + V$. Find $\text{cov}(R, S)$

The covariance between two random variables R and S is defined as:

$$\text{cov}(R, S) = E[(R - E[R])(S - E[S])] = E[RS] - E[R]E[S]$$

First, let's find the expected values of R and S:

$$\bullet E[R] = E[5 + 2U - 3UV] = 5 + 2E[U] - 3E[UV]$$

Since U and V are independent, $E[UV] = E[U]E[V]$. Also, $E[U] = E[V] = 0$, as U and V are standard normal.

$$\text{Therefore, } E[R] = 5 + 2(0) - 3(0)(0) = 5$$

$$\bullet E[S] = E[2 - U + V] = 2 - E[U] + E[V] = 2 - 0 + 0 = 2$$

Now, let's find $E[RS]$:

$$\begin{aligned} E[RS] &= E[(5 + 2U - 3UV)(2 - U + V)] = E[10 - 5U + 5V + 4U - 2U^2 + 2UV - 6UV + 3U^2V - 3UV^2] \\ &= 10 - 5E[U] + 5E[V] + 4E[U] - 2E[U^2] + 2E[UV] - 6E[UV] + 3E[U^2V] - 3E[UV^2] \end{aligned}$$

Since U and V are standard normal, $E[U] = E[V] = 0$ and $E[U^2] = E[V^2] = 1$.

Also, since U and V are independent, $E[UV] = E[U]E[V] = 0$, $E[U^2V] = E[U^2]E[V] = 1 * 0 = 0$, and $E[UV^2] = E[U]E[V^2] = 0 * 1 = 0$.

$$\text{Therefore, } E[RS] = 10 - 5(0) + 5(0) + 4(0) - 2(1) + 2(0) - 6(0) + 3(0) - 3(0) = 10 - 2 = 8$$

$$\text{Finally, we can find the covariance: } \text{cov}(R, S) = E[RS] - E[R]E[S] = 8 - (5)(2) = 8 - 10 = -2$$

Probability Theory

Events and Sample Space

- Sample Space:
 - for a procedure Sample Space consists of all possible simple events; that is, the sample space consists of all outcomes that cannot be broken down any further
- Event
 - any collection of results or outcomes of a procedure
- Simple Event
 - an outcome or an event that cannot be further broken down into simpler components
 - Sample space Ω - set of all possible outcomes of a random experiment
 - Dice roll: {1, 2, 3, 4, 5, 6}
 - Coin toss: {Tails, Heads}
 - Event space \mathcal{F} - subsets of elements in a sample space
 - Dice roll: {1, 2, 3} or {2, 4, 6}
 - Coin toss: {Tails}

Events and Sample Space

- A pair of dice are rolled. The sample space has 36 simple events:

1,1 1,2 1,3 1,4 1,5 1,6

2,1 2,2 2,3 2,4 2,5 2,6

3,1 3,2 3,3 3,4 3,5 3,6

4,1 4,2 4,3 4,4 4,5 4,6

5,1 5,2 5,3 5,4 5,5 5,6

6,1 6,2 6,3 6,4 6,5 6,6

where the pairs represent the numbers rolled on each dice.

- Which elements of the sample space correspond to the event that the sum of each dice is 4?

Probability

- The word 'Probability' means the chance of occurring of a particular event.
- It is generally possible to predict the future of an event quantitatively with a certain probability of being correct.
- The probability is used in such cases where the outcome of the trial is uncertain.

$$P(A) = \frac{\text{number of cases favourable to } A}{\text{number of possible outcomes}}$$

- P - denotes a probability.
- A, B, and C - denote specific events.
- $P(A)$ - denotes the probability of event A occurring.

Probability

- Probability of an Event Defined over (Ω, \mathcal{F}) s.t.
 - $0 < P(a) < 1$ for all a in \mathcal{F}
 - $P(\Omega) = 1$
- Probability of an event which is certain to occur is **one**.
- Probability of an event which is impossible to **zero**.
- If the probability of happening of an event $P(A)$ and that of not happening is $P(A')$, then $P(A) + P(A') = 1$,
where, $0 \leq P(A) \leq 1$, $0 \leq P(A') \leq 1$.

Event Relations

- **Equally Likely Events:** Events are said to be equally likely if one of them cannot be expected to occur in preference to others. In other words, it means each outcome is as likely to occur as any other outcome.
 - *Example:* When a die is thrown, all the six faces, i.e., 1, 2, 3, 4, 5 and 6 are equally likely to occur.
- **Mutually Exclusive or Disjoint Events:** Events are called mutually exclusive if they cannot occur simultaneously.
 - *Example:* Suppose a card is drawn from a pack of cards, then the events getting a jack and getting a king are mutually exclusive because they cannot occur simultaneously.

Event Relations

- **Exhaustive Events:** The total number of all possible outcomes of an experiment is called exhaustive events.
 - Example: In the tossing of a coin, either head or tail may turn up. Therefore, there are two possible outcomes. Hence, there are two exhaustive events in tossing a coin.
- **Dependent Event:** Events are said to be dependent if occurrence of one affect the occurrence of other events.
- **Independent Events:** Events A and B are said to be independent if the occurrence of any one event does not affect the occurrence of any other event.

$$P(A \cap B) = P(A) P(B).$$

Event Relations

Example: A coin is tossed thrice, and all 8 outcomes are equally likely

A: "The first throw results in heads."

B: "The last throw results in Tails."

Prove that event A and B are independent.

Solution:

Sample Space: [HHH, HHT, HTH, THH, TTT, TTH, THT, HTT]

A: [HHH, HHT, HTH, HTT]

B: [HHT, TTT, THT, HTT]

AnB: [HHT, HTT]

$$P(A) = \frac{4}{8} = \frac{1}{2}$$

$$P(B) = \frac{4}{8} = \frac{1}{2}$$

$$P(AnB) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Event Relations

Theorem 1: If A and B are two mutually exclusive events, then

$$P(A \cup B) = P(A) + P(B)$$

Proof: Let the n=total number of exhaustive cases

n_1 = number of cases favorable to A.

n_2 = number of cases favorable to B.

Now, we have A and B two mutually exclusive events. Therefore, $n_1 + n_2$ is the number of cases favorable to A or B.

$$P(A \cup B) = \frac{\text{favorable cases}}{\text{Total number of exhaustive cases}} = \frac{n_1 + n_2}{n} = \frac{n_1}{n} + \frac{n_2}{n}$$

But we have, $P(A) = \frac{n_1}{n}$ and $P(B) = \frac{n_2}{n}$

Hence, $P(A \cup B) = P(A) + P(B)$.

Event Relations

Theorem2: If A and B are two events that are not mutually exclusive, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Proof: Let n = total number of exhaustive cases

n_1 =number of cases favorable to A

n_2 = number of cases favorable to B

n_3 = number of cases favorable to both A and B

But A and B are not mutually exclusive. Therefore, A and B can occur simultaneously. So, $n_1+n_2-n_3$ is the number of cases favorable to A or B.

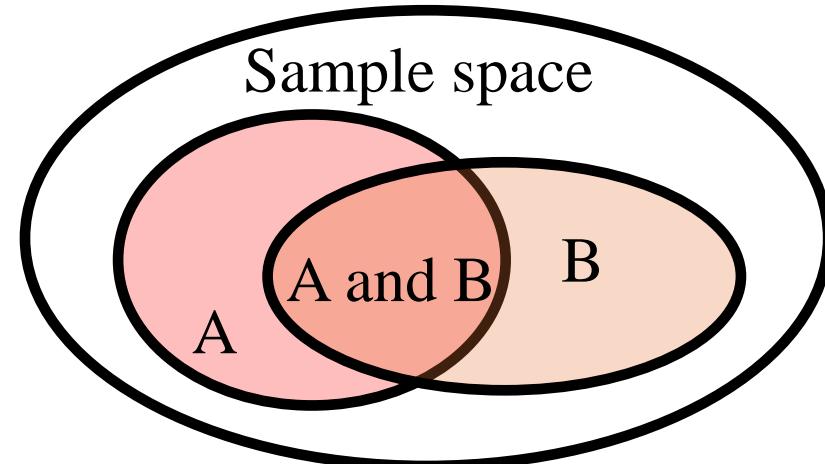
$$\text{Therefore, } P(A \cup B) = \frac{n_1 + n_2 - n_3}{n} = \frac{n_1}{n} + \frac{n_2}{n} - \frac{n_3}{n}$$

$$\text{But we have, } P(A) = \frac{n_1}{n}, P(B) = \frac{n_2}{n} \text{ and } P(A \cap B) = \frac{n_3}{n}$$

$$\text{Hence, } P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Conditional Probability

- The probability of an event A based on the occurrence of another event B is termed conditional Probability. It is denoted as $P(A|B)$ and represents the probability of A when event B has already happened.
- $P(A | B) = P(A \cap B) / P(B)$
- If the two events are independent:
 - $P(A \cap B) = P(A) * P(B)$
 - $P(A/B) = P(A)$



Joint Probability:

The probability of two more events occurring together and at the same time is measured it is termed as Joint Probability.

Joint probability for two events A and B is denoted as, $P(A \cap B)$.

Bayes Theorem

- Bayes theorem, also known as the Bayes Rule, is used to determine the conditional probability of event A when event B has already happened.
- “The conditional probability of an event A, given the occurrence of another event B, is equal to the product of the probability event of B given A and the probability of A divided by the probability of event B.” i.e.

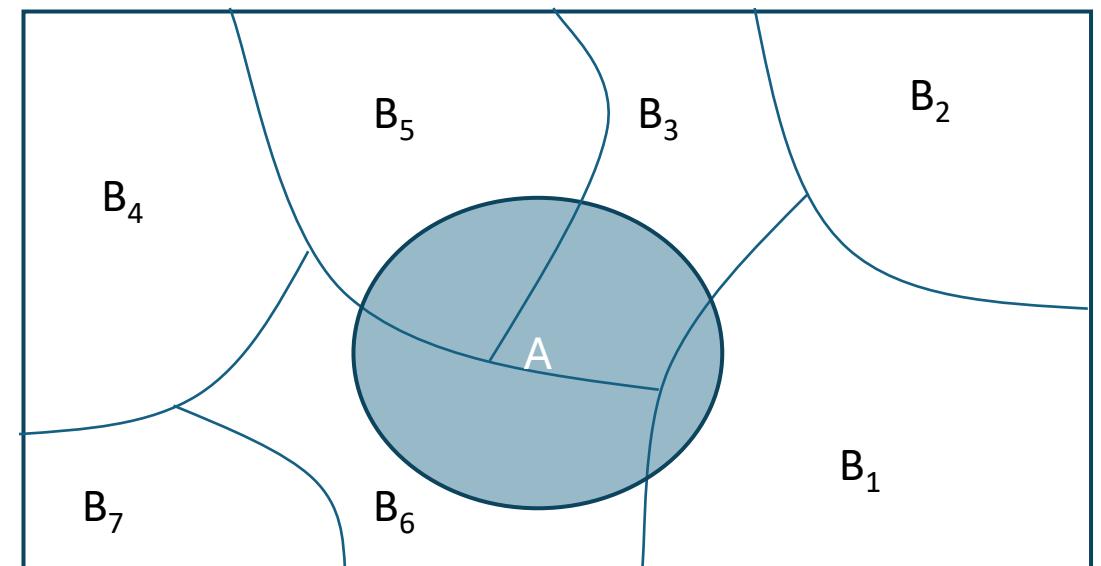
$$P(A|B) = P(B|A)P(A) / P(B) \text{ given } P(B) \neq 0$$

- where,
 - $P(A)$ and $P(B)$ are the probabilities of events A and B
 - $P(A|B)$ is the probability of event A when event B happens
 - $P(B|A)$ is the probability of event B when event A happens

Theorem of Total Probability

- Let E_1, E_2, \dots, E_n are mutually exclusive and exhaustive events associated with a random experiment and let E be an event that occurs with some E_i .
- Then,

$$P(E) = \sum_{i=1}^n P(E|E_i) \cdot P(E_i)$$



$$p(A) = \sum P(B_i)P(A|B_i)$$

Questions

Q1. Two dice are thrown. The events A, B, C, D, E, F

A = getting even number on first die.

B= getting an odd number on the first die.

C = getting a sum of the number on dice ≤ 5

D = getting a sum of the number on dice > 5 but less than 10.

Show that:

1. A, B are a mutually exclusive event and Exhaustive Event.
2. A, C are not mutually exclusive.
3. C, D are a mutually exclusive event but not Exhaustive Event.
4. A' and B' are a mutually exclusive and exhaustive event.

Questions

Q2. A bag contains 5 green and 7 red balls. Two balls are drawn. Find the probability that one is green and the other is red.

Q3. Find the probability of drawing a heart on each of two consecutive draws from well shuffled-packs of cards if the card is not replaced after the draw.

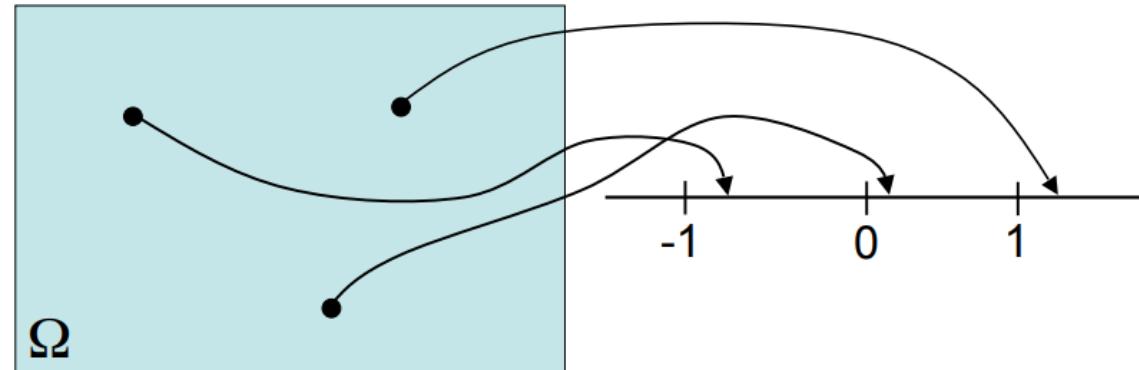
Q4. “ $X+Y=6$ or $X+Y=7$ ” – given this (and only this), what is the probability of $Y=5$?

Q5. There are three urns containing 3 white and 2 black balls; 2 white and 3 black balls; 1 black and 4 white balls respectively. There is an equal probability of each urn being chosen. One ball is equal probability chosen at random. what is the probability that a white ball is drawn?

Probability Theory

Random Variable

- A random variable is a numerical quantity that is generated by a random experiment.
- A RV is any rule (i.e., function) that associates a number with each outcome in the sample space.



Example 1 : Machine Breakdowns

- Sample space : $S = \{\text{electrical, mechanical, misuse}\}$
- Each of these failures may be associated with a repair cost
- State space : {50, 200, 350}
- Cost is a random variable : 50, 200, and 350

Random Variable

- We will denote random variables by capital letters, such as X or Z , and the actual values that they can take by lowercase letters, such as x and z .
- A RV is called **discrete** if its possible values form a finite or countable set.
- A RV is called **continuous** if its possible values contain a whole interval of numbers.

Experiment	Number X	Possible Values of X	Type of RV
Roll two fair dice	Sum of the number of dots on the top faces	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12	discrete
Flip a fair coin repeatedly	Number of tosses until the coin lands heads	1, 2, 3, 4, ...	discrete
Measure the voltage at an electrical outlet	Voltage measured	$118 \leq x \leq 122$	continuous
Operate a light bulb until it burns out	Time until the bulb burns out	$0 \leq x < \infty$	continuous

Probability Distribution

- The probability distribution for a random variable describes how the probabilities are distributed over the values of the random variable.
- The probabilities of a RV X must satisfy the following two conditions:
 - Each probability $P(x)$ must be between 0 to 1: $0 \leq P(x) \leq 1$.
 - The sum of all the possible probabilities is 1: $\sum P(x) = 1$.
- For a **discrete random variable**, x , the probability distribution is defined by a **probability mass function**, denoted by $f(x)$.
- This function provides the probability for each value of the random variable.

Probability Mass Function (p.m.f.)

- A set of probability value p_i assigned to each of the values taken by the discrete random variable x_i
- $0 \leq p_i \leq 1$ and $\sum_i p_i = 1$
- Probability : $P(X = x_i) = p_i$

Probability Mass Function

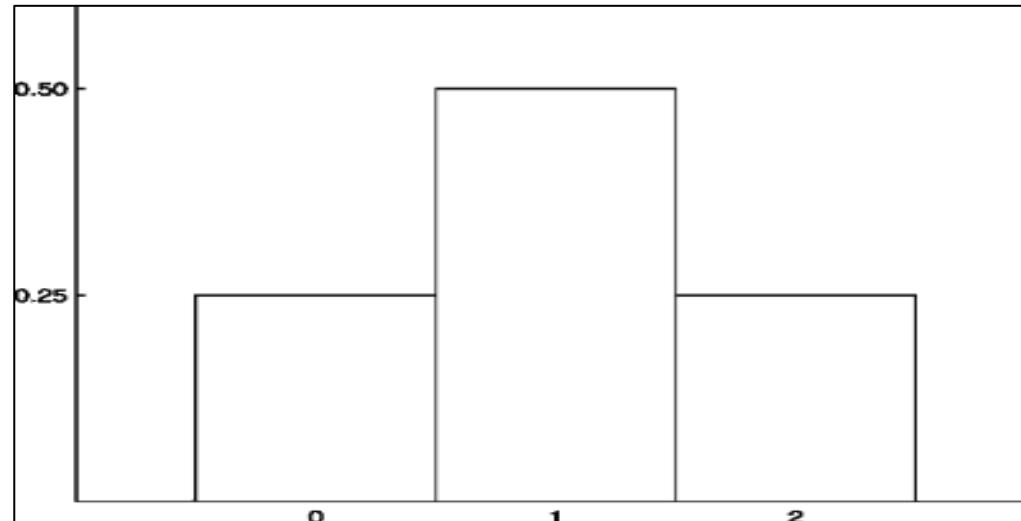
Q1. A fair coin is tossed twice. Let X be the number of heads that are observed.

- Construct the probability distribution of X .
- Find the probability that at least one head is observed.

The possible values that X can take are 0, 1, and 2.

- $S=\{hh,ht,th,tt\}$ are equally likely outcomes
 - $X=0$ to $\{tt\}$, $X=1$ to $\{ht,th\}$, and $X=2$ to $\{hh\}$.
- The probability distribution of X , is given by

x	0	1	2
$P(x)$	0.25	0.50	0.25



“At least one head” is the event $X \geq 1$, which is the union of the mutually exclusive events $X=1$ and $X=2$.

- Thus, $P(X \geq 1) = P(1) + P(2) = 0.50 + 0.25 = 0.75$

Probability Mass Function

Q2: A pair of fair dice is rolled. Let X denote the sum of the number of dots on the top faces.

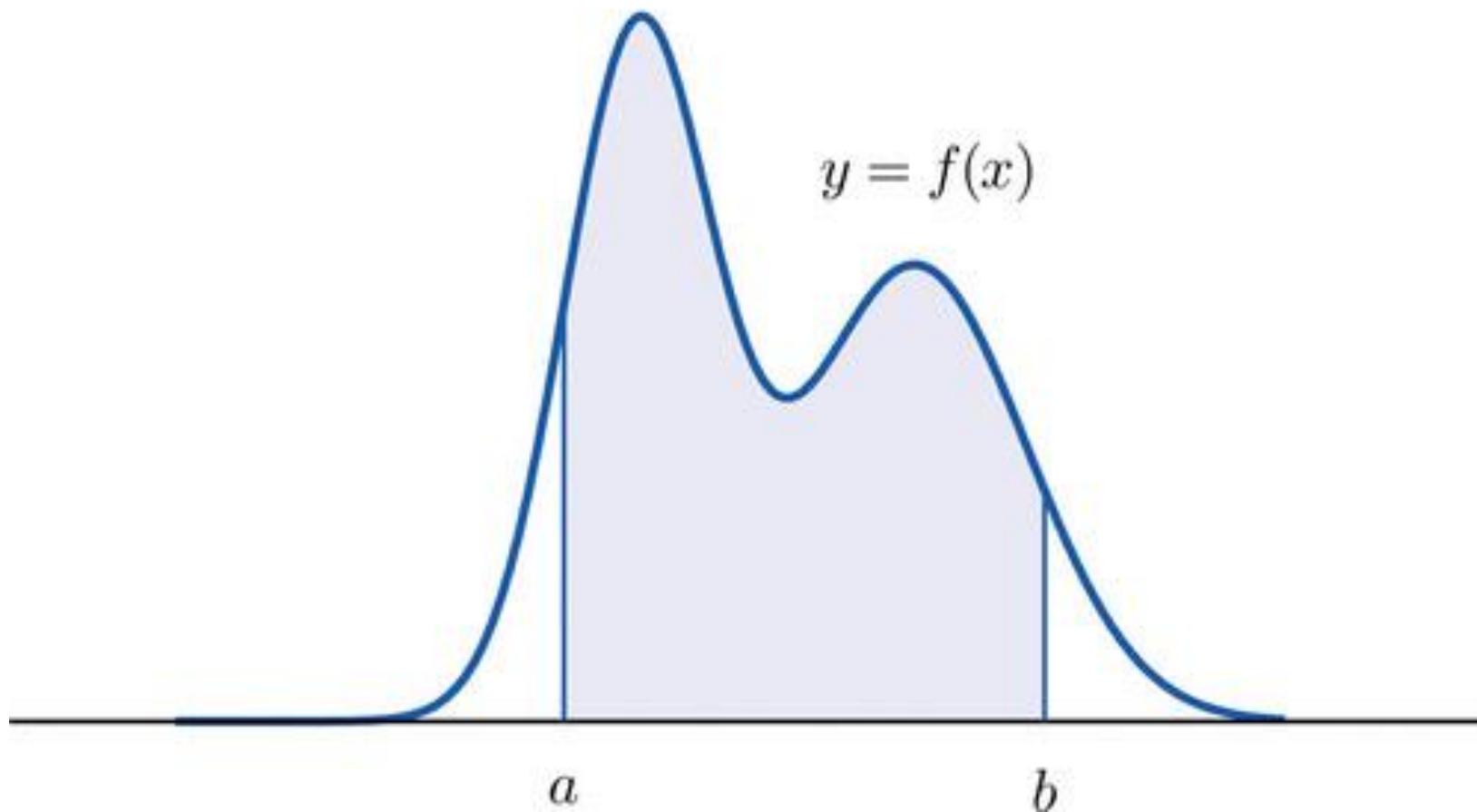
- Construct probability distribution of X for a pair of fair dice.
- Find $P(X \geq 9)$ Ans: 10/36
- Find the probability that X takes an even value. Ans: 0.5

Probability Density Function

- The probability distribution of a continuous random variable X is an assignment of probabilities to intervals of decimal numbers using a function $f(x)$, called a **Probability Density Function**.
- The probability that X assumes a value in interval $[a,b]$ is equal to the area of the region that is bounded above by the graph of the equation $y=f(x)$, bounded below by the x -axis, and bounded on the left and right by the vertical lines through a and b .
- Density Function $f(x)$ must satisfy the following two conditions:
 - For all numbers x , $f(x) \geq 0$, so that the graph of $y=f(x)$ never drops below the x -axis.
 - The area of the region under the graph of $y=f(x)$ and above the x -axis is 1.

Probability Density Function

$P(a < X < b) = \text{area of shaded region}$



Probability Density Function

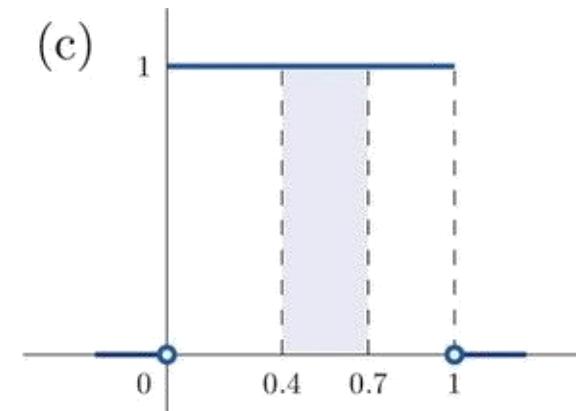
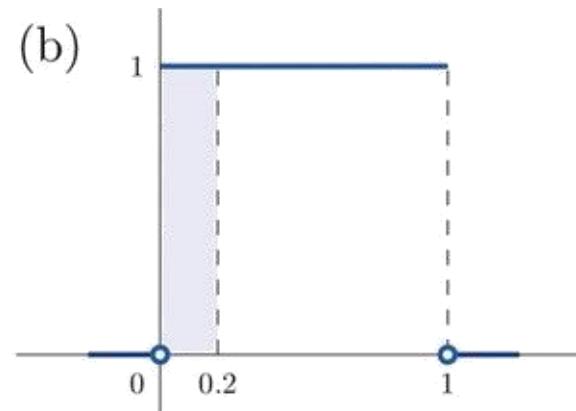
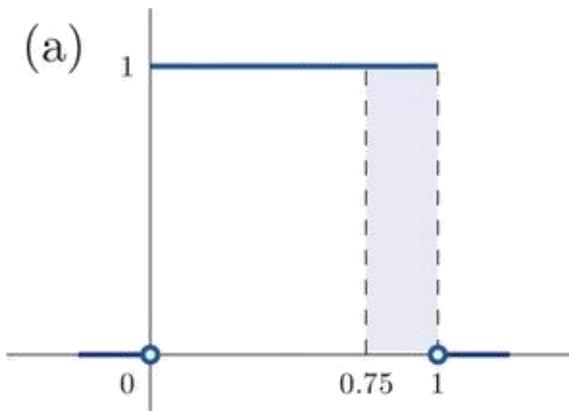
Q3. A random variable X has the uniform distribution on the interval $[0,1]$: the density function is $f(x)=1$ if x is between 0 and 1 and $f(x)=0$ for all other values of x (a uniform distribution).

- Find $P(X>0.75)$, the probability that X assumes a value greater than 0.75
- Find $P(X\leq 0.2)$, the probability that X assumes a value less than or equal to 0.2
- Find $P(0.4 < X < 0.7)$, the probability that X assumes a value between 0.4 and 0.7

Q4. A man arrives at a bus stop at a random time (that is, with no regard for the scheduled service) to catch the next bus. Buses run every 30 minutes without fail, hence the next bus will come any time during the next 30 minutes with evenly distributed probability (a uniform distribution). Find the probability that a bus will come within the next 10 minutes.

Probability Density Function

- $P(X>0.75)$ is the area of the rectangle of height 1 and base length $1-0.75=0.25$, hence is $\text{base} \times \text{height} = (0.25) \cdot (1) = 0.25$
- $P(X \leq 0.2)$ is the area of the rectangle of height 1 and base length $0.2-0=0.2$, hence is $\text{base} \times \text{height} = (0.2) \cdot (1) = 0.2$
- $P(0.4 < X < 0.7)$ is the area of the rectangle of height 1 and length $0.7-0.4=0.3$, hence is $\text{base} \times \text{height} = (0.3) \cdot (1) = 0.3$



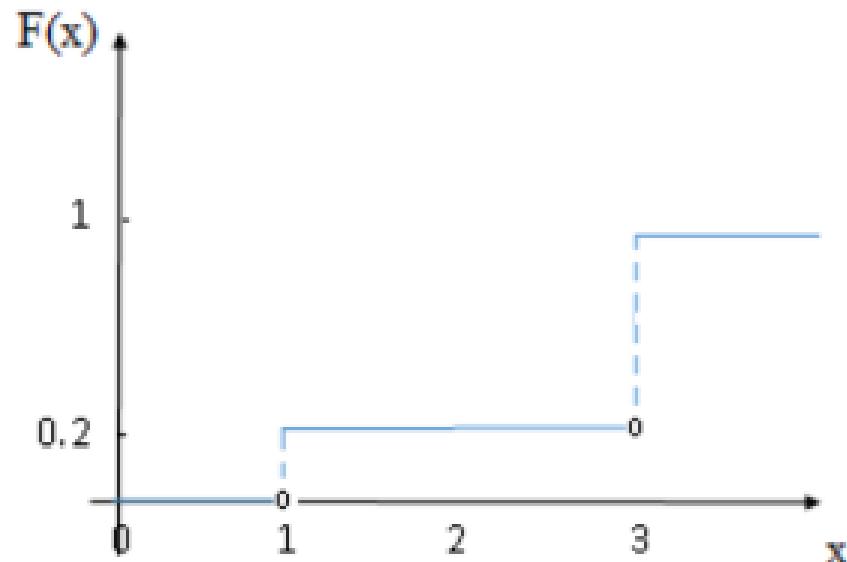
CDF of Random Variable

- Suppose that X is a random variable with values in \mathbb{R} . The Cumulative Distribution Function of X is the function $F:\mathbb{R}\rightarrow[0,1]$ is defined by

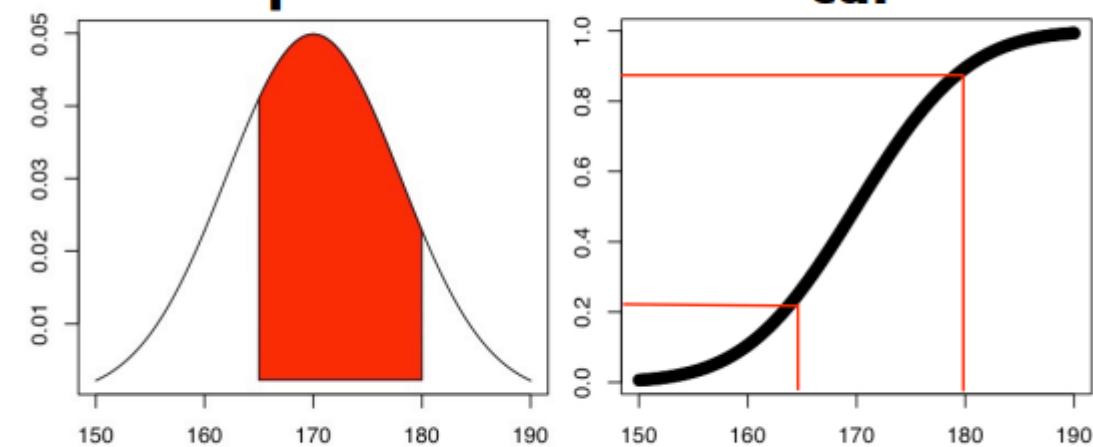
$$F(x) = P(X \leq x), \quad x \in \mathbb{R}$$

- Discrete RV**

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} p(x_i),$$



Continuous rv: $F(x) = P(X \leq x) = \int_{-\infty}^x f(y)dy$



$$P(a \leq X \leq b) = F(b) - F(a)$$

Probability Theory

Expectation of Random Variable

- The expected value, or mean, of a random variable x denoted by $E(x)$ or μ is a weighted average of the values the random variable may assume.
- In the discrete case the weights are given by the probability mass function, and in the continuous case the weights are given by the probability density function.
- The expectation of a discrete random variable X is given by:
$$\mu=E(X) = \sum xf(x),$$
 where $f(x)$ is the PMF of $x.$
- The expectation of a continuous random variable X is given by:
$$\mu=E(X)=\int xf(x)dx,$$
 where $f(x)$ is the PDF of $x.$

Variance of Random Variable

- The **variance** of a random variable, denoted by $\text{Var}(x)$ or σ^2 , is a weighted average of the squared deviations from the mean.
 - For discrete RV: $\text{Var}(x) = \sigma^2 = \sum(x - \mu)^2f(x)$
 - For continuous RV: $\text{Var}(x) = \sigma^2 = \int(x - \mu)^2f(x)dx$
- The **standard deviation**, denoted σ , is the positive square root of the variance.

Median of Random Variable

- The **median** of the discrete random variable X , is the value of x for which $P(X \leq x)$ is greater than or equal to 0.5 and $P(X \geq x)$ is greater than or equal to 0.5.

Median of Random Variable

- Let X be a continuous rv with probability density function, $f(x)$. The median of X can be obtained by solving for c in the equation below:

$$\int_{-\infty}^c f(x)dx = 0.5$$

- That is, it is the value for which the area under the curve from negative infinity to c is equal to 0.50.

Quantiles of Random Variable

- Let X is a real-valued random variable with Cumulative Distribution Function F .
- For $p \in (0,1)$, a value of x is called **a quantile of order p** for the distribution if
$$F(x-) = P(X < x) \leq p \text{ and } F(x) = P(X \leq x) \geq p .$$
- A quantile of order p is a value where the graph of the cumulative distribution function crosses p .
- Median is also called 50^{th} percentile.

Questions

Q1. A service organization in a large town organizes a raffle each month. One thousand raffle tickets are sold for \$1 each. Each has an equal chance of winning. First prize is \$300, second prize is \$200, and third prize is \$100. Let, X denote the net gain from the purchase of one ticket.

- Construct the probability distribution of X
- Find the probability of winning any money in the purchase of one ticket.
- Find the expected value of X , and interpret its meaning.

Questions

a) If a ticket is selected as the first prize winner, the net gain to the purchaser is the \$300 prize less the \$1 that was paid for the ticket, hence $X=300-11=299$. There is one such ticket, so $P(299)=0.001$

Applying the same “income minus outgo” principle to the second and third prize winners and to the 997 losing tickets yields the probability distribution:

x	299	199	99	-1
$P(x)$	0.001	0.001	0.001	0.997

b) Let W denote the event that a ticket is selected to win one of the prizes. Using the table

$$P(W) = P(299)+P(199)+P(99) = 0.003$$

c) $E(X) = (299)\cdot(0.001) + (199)\cdot(0.001) + (99)\cdot(0.001) + (-1)\cdot(0.997)$
 $= -0.4$

Questions

Q2. A discrete rv X has following probability distribution:

x	-1	0	1	4
P(x)	0.2	0.5	c	0.1

Find

- a) c 0.2
- b) P(0) 0.5
- c) P($X > 0$) 0.3
- d) P($X \geq 0$) 0.8
- e) P($X \leq -2$) 0
- f) The mean μ of X 0.4
- g) The variance σ^2 of X 1.84
- h) The standard deviation σ of X 1.3565

Questions

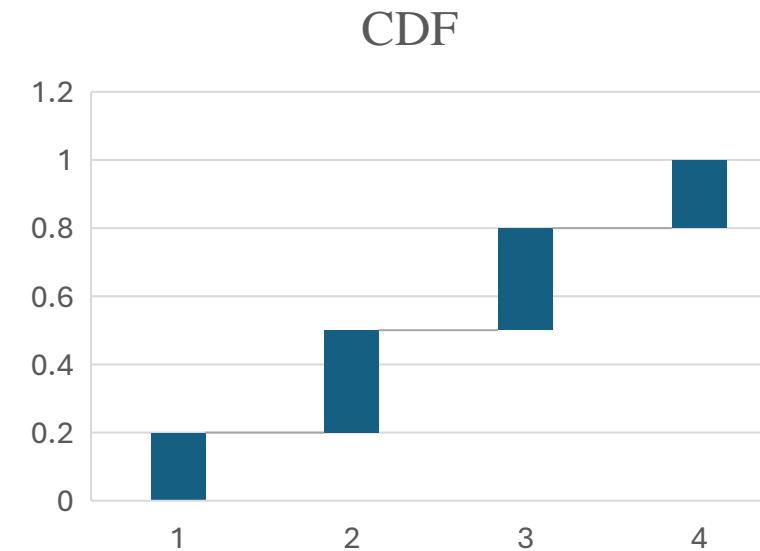
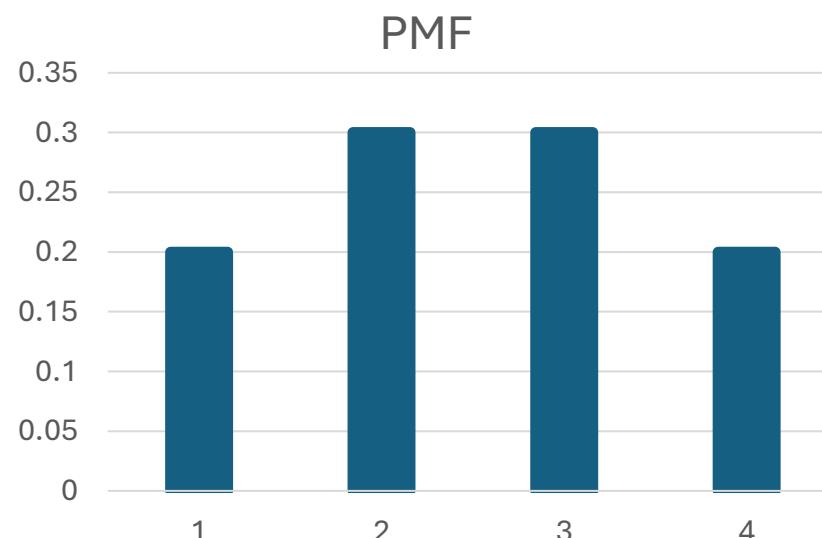
Q3. Given the following probability density function of a discrete random variable,

calculate the median of the distribution: $f(x) = \begin{cases} 0.2 & x = 1, 4 \\ 0.3 & x = 2, 3 \end{cases}$

- Solution: The median of the distribution above is 2 because;

$$P(X \leq 2) = P(X=1) + P(X=2) = 0.2 + 0.3 = 0.5 \text{ and,}$$

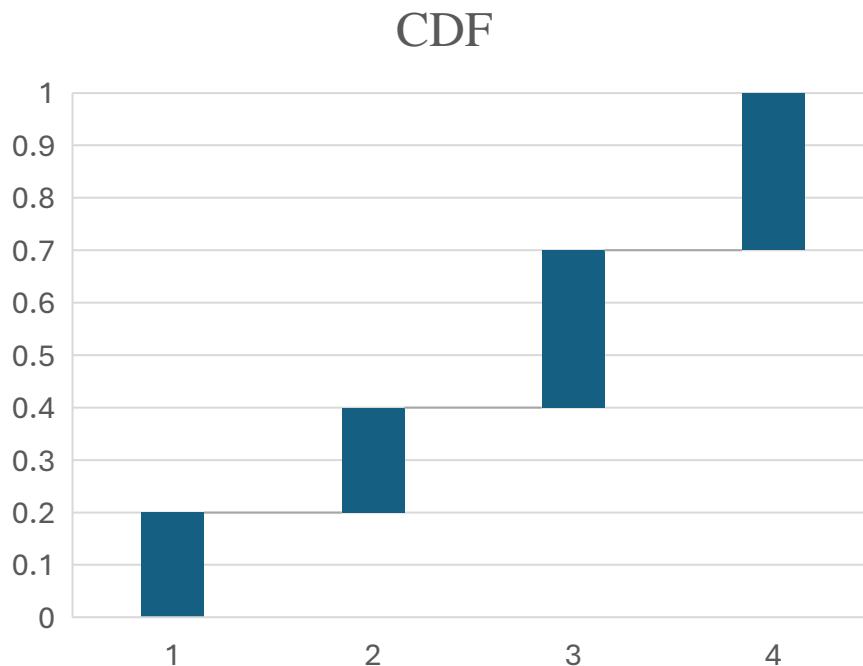
$$P(X \geq 2) = P(X=2) + P(X=3) + P(X=4) = 0.3 + 0.3 + 0.2 = 0.8$$



Questions

Q4. Given the following probability density function of a discrete random variable, calculate the 75th Percentile of the distribution:

$$f(x) = \begin{cases} 0.2 & x = 1, 4 \\ 0.3 & x = 2, 3 \end{cases}$$



Solution: The 75th percentile of the distribution is 4 because;

$$\begin{aligned} P(X < 3) &= P(X=1)+P(X=2) \\ &= 0.2+0.3 = 0.5 \text{ and,} \end{aligned}$$

$$\begin{aligned} P(X \leq 3) &= P(X=1)+P(X=2)+P(X=3) \\ &= 0.2+0.3+0.3=0.8 \end{aligned}$$

Probability Theory

Joint Distribution of RVs

- In real life, we are often interested in two (or more) random variables at the same time. For example,
 - we might measure the height and weight of an object, or
 - frequency of exercise and rate of heart disease in adults,
 - level of air pollution and rate of respiratory illness in cities,
 - number of Facebook friends and age of Facebook members
- Joint distribution allows us to compute probabilities of events involving both variables and understand the relationship between the variables.

Joint Distribution of Discrete RVs

- Suppose X and Y are two discrete random variables.
 - X takes values $\{x_1, x_2, \dots, x_n\}$ and Y takes values $\{y_1, y_2, \dots, y_m\}$. The ordered pair (X, Y) take values in the product $\{(x_1, y_1), (x_1, y_2), \dots, (x_n, y_m)\}$.
 - The joint probability mass function (joint pmf) of X and Y is the function $p(x_i, y_j)$ giving the probability of the joint outcome $X = x_i, Y = y_j$.

Joint probability mass function must satisfy two properties:

1. $0 \leq p(x_i, y_j) \leq 1$
2. The total probability is 1.

$$\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) = 1$$

$X \setminus Y$	y_1	y_2	\dots	y_j	\dots	y_m
x_1	$p(x_1, y_1)$	$p(x_1, y_2)$	\dots	$p(x_1, y_j)$	\dots	$p(x_1, y_m)$
x_2	$p(x_2, y_1)$	$p(x_2, y_2)$	\dots	$p(x_2, y_j)$	\dots	$p(x_2, y_m)$
\dots	\dots	\dots	\dots	\dots	\dots	\dots
\dots	\dots	\dots	\dots	\dots	\dots	\dots
x_i	$p(x_i, y_1)$	$p(x_i, y_2)$	\dots	$p(x_i, y_j)$	\dots	$p(x_i, y_m)$
\dots	\dots	\dots	\dots	\dots	\dots	\dots
x_n	$p(x_n, y_1)$	$p(x_n, y_2)$	\dots	$p(x_n, y_j)$	\dots	$p(x_n, y_m)$

Joint Distribution of Discrete RVs

Q1. Roll two dice. Let X be the value on the first die and let T be the total on both dice. Draw the joint probability table.

$X \setminus T$	2	3	4	5	6	7	8	9	10	11	12
1	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	0
2	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0
3	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0
4	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0
5	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0
6	0	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36

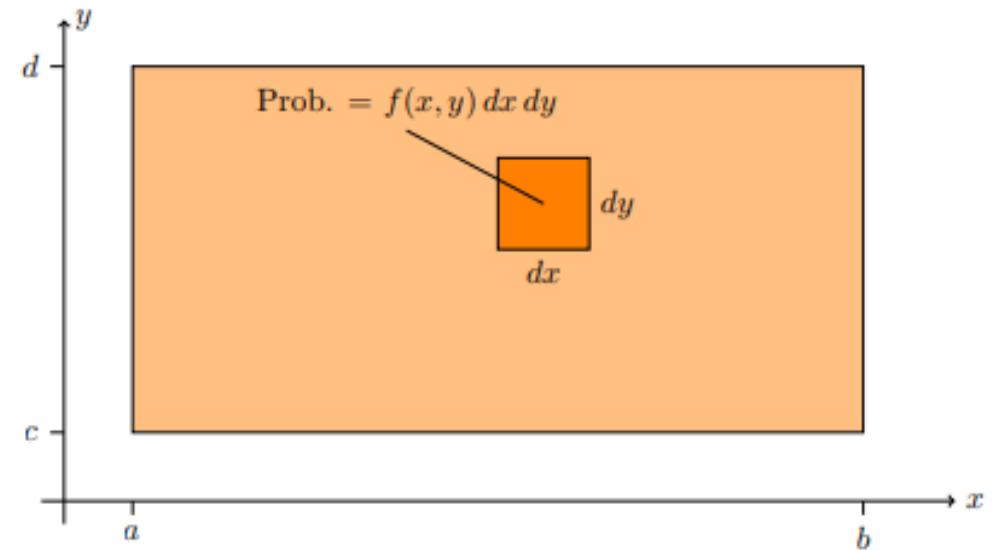
Q2. Roll two dice. Let X be the value on the first die and let Y be the value on the second die. Then both X and Y take values 1 to 6 and the joint pmf is $p(i, j) = 1/36$ for all i and j between 1 and 6. Draw the Joint probability table and find the probability of event B= ' $X-Y \geq 2$ '.

Joint Distribution of Continuous RVs

If X takes values in $[a, b]$ and Y takes values in $[c, d]$ then the pair (X, Y) takes values in the product $[a, b] \times [c, d]$.

- The joint probability density function (joint pdf) of X and Y is a function $f(x, y)$ giving the probability density at (x, y) .
- That is, the probability that (X, Y) is in a small rectangle of width dx and height dy around (x, y) is $f(x, y)dx dy$.
- A joint PDF must satisfy:
 1. $0 \leq f(x, y)$
 2. The total probability is 1.

$$\int_c^d \int_a^b f(x, y) dx dy = 1$$



Joint Cumulative Distributions RVs

Suppose X and Y are jointly-distributed random variables. We will use the notation ' $X \leq x, Y \leq y$ ' to mean the event ' $X \leq x$ and $Y \leq y$ '. The **joint cumulative distribution function** (joint cdf) is defined as

$$F(x, y) = P(X \leq x, Y \leq y)$$

Continuous case: If X and Y are continuous random variables with joint density $f(x, y)$ over the range $[a, b] \times [c, d]$ then the joint cdf is given by the double integral

$$F(x, y) = \int_c^y \int_a^x f(u, v) du dv.$$

To recover the joint pdf, we differentiate the joint cdf. Because there are two variables we need to use partial derivatives:

$$f(x, y) = \frac{\partial^2 F}{\partial x \partial y}(x, y).$$

Discrete case: If X and Y are discrete random variables with joint pmf $p(x_i, y_j)$ then the joint cdf is give by the double sum

$$F(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} p(x_i, y_j).$$

Joint Distribution of Continuous RVs

Q3. Let X & Y both take values in [0,1] with density $f(x, y) = 4xy$.

- i. Show $f(x, y)$ is a valid joint PDF,
- ii. Visualize the event $A = 'X < 0.5 \text{ and } Y > 0.5'$ and find its probability.

To show $f(x, y)$ is a valid joint pdf we must check that it is positive (which it clearly is) and that the total probability is 1.

$$\text{Total probability} = \int_0^1 \int_0^1 4xy \, dx \, dy = \int_0^1 [2x^2y]_0^1 \, dy = \int_0^1 2y \, dy = 1. \quad \text{QED}$$

The event A is just the upper-left-hand quadrant. Because the density is not constant we must compute an integral to find the probability.

$$P(A) = \int_0^{.5} \int_{.5}^1 4xy \, dy \, dx = \int_0^{.5} [2xy^2]_{.5}^1 \, dx = \int_0^{.5} \frac{3x}{2} \, dx = \boxed{\frac{3}{16}}.$$

Q4. Let X & Y both take values in [0,1] with density $f(x, y) = 4xy$. Find Joint CDF of X and Y.

Marginal Density RVs

Given a joint density for X and Y , we define the marginal density of X to be

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

and the marginal density of Y to be

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

As usual, we restrict the integral to the region where f is positive when that is not the entire plane.

Example Consider

$$f(x, y) = \begin{cases} x + y & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The marginal density of X is given by

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x, y) dy \\ &= \int_0^1 x + y dy \\ &= x + 1/2 \end{aligned}$$

Marginal Distributions RVs

Q5. Suppose (X, Y) takes values on the unit square $[0, 1] \times [0, 1]$ with joint pdf $f(x, y) = \frac{3}{2} (x^2 + y^2)$. Find the marginal pdf $f_X(x)$ and use it to find $P(X < 0.5)$.

$$f_X(x) = \int_0^1 \frac{3}{2} (x^2 + y^2) dy = \left[\frac{3}{2} x^2 y + \frac{y^3}{2} \right]_0^1 = \boxed{\frac{3}{2} x^2 + \frac{1}{2}}.$$

$$P(X < 0.5) = \int_0^{0.5} f_X(x) dx = \int_0^{0.5} \frac{3}{2} x^2 + \frac{1}{2} dx = \left[\frac{1}{2} x^3 + \frac{1}{2} x \right]_0^{0.5} = \boxed{\frac{5}{16}}.$$

Independence in RVs

- Events A and B are independent if $P(A \cap B) = P(A)P(B)$.
- The joint distribution (or density or mass) of Independent RVs is the product of the marginals.

Definition: Jointly-distributed random variables X and Y are **independent** if their joint cdf is the product of the marginal cdf's:

$$F(X, Y) = F_X(x)F_Y(y).$$

For discrete variables this is equivalent to the joint pmf being the product of the marginal pmf's.:

$$p(x_i, y_j) = p_X(x_i)p_Y(y_j).$$

For continuous variables this is equivalent to the joint pdf being the product of the marginal pdf's.:

$$f(x, y) = f_X(x)f_Y(y).$$

Independence in RVs

Example 12. For discrete variables independence means the probability in a cell must be the product of the marginal probabilities of its row and column. In the first table below this is true: every marginal probability is $1/6$ and every cell contains $1/36$, i.e. the product of the marginals. Therefore X and Y are independent.

$X \setminus Y$	1	2	3	4	5	6	$p(x_i)$
1	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/6$
2	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/6$
3	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/6$
4	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/6$
5	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/6$
6	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/6$
$p(y_j)$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	1

Example 13. For continuous variables independence means you can factor the joint pdf or cdf as the product of a function of x and a function of y .

- (i) Suppose X has range $[0, 1/2]$, Y has range $[0, 1]$ and $f(x, y) = 96x^2y^3$ then X and Y are independent. The marginal densities are $f_X(x) = 24x^2$ and $f_Y(y) = 4y^3$.
- (ii) If $f(x, y) = 1.5(x^2 + y^2)$ over the unit square then X and Y are not independent because there is no way to factor $f(x, y)$ into a product $f_X(x)f_Y(y)$.
- (iii) If $F(x, y) = \frac{1}{2}(x^3y + xy^3)$ over the unit square then X and Y are not independent because the cdf does not factor into a product $F_X(x)F_Y(y)$.

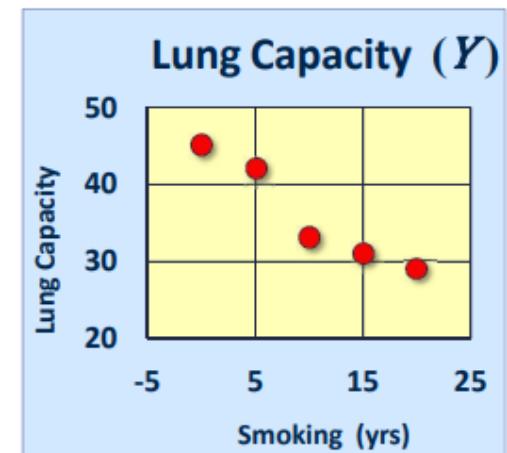
Probability Theory

Covariance of RVs

- Random Variables may change in relation to each other. Covariance is a measure of association of two variables.
- If positive, then both variables increase or decrease together. If negative, then they vary in opposite manner.
- Covariance measures how much the movement in one variable predicts the movement in a corresponding variable
- **Example:** investigate relationship between cigarette smoking and lung capacity as shown in figure.

N	Cigarettes (X)	Lung Capacity (Y)
1	0	45
2	5	42
3	10	33
4	15	31
5	20	29

- Variables smoking and lung capacity covary inversely, like



Covariance of RVs

- Average product of deviation measures extent to which variables co-vary, the degree of linkage between them.

$$S_{xy} = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$$

↑ ↑
Deviation of data 1 Deviation from
from mean mean of data 2

Cigs (X)				Cap (Y)
0	-10	-90	9	45
5	-5	-30	6	42
10	0	0	-3	33
15	5	-25	-5	31
20	10	-70	-7	29
$\Sigma = -215$				

Evaluation yields,

$$S_{xy} = \frac{1}{4}(-215) = -53.75$$

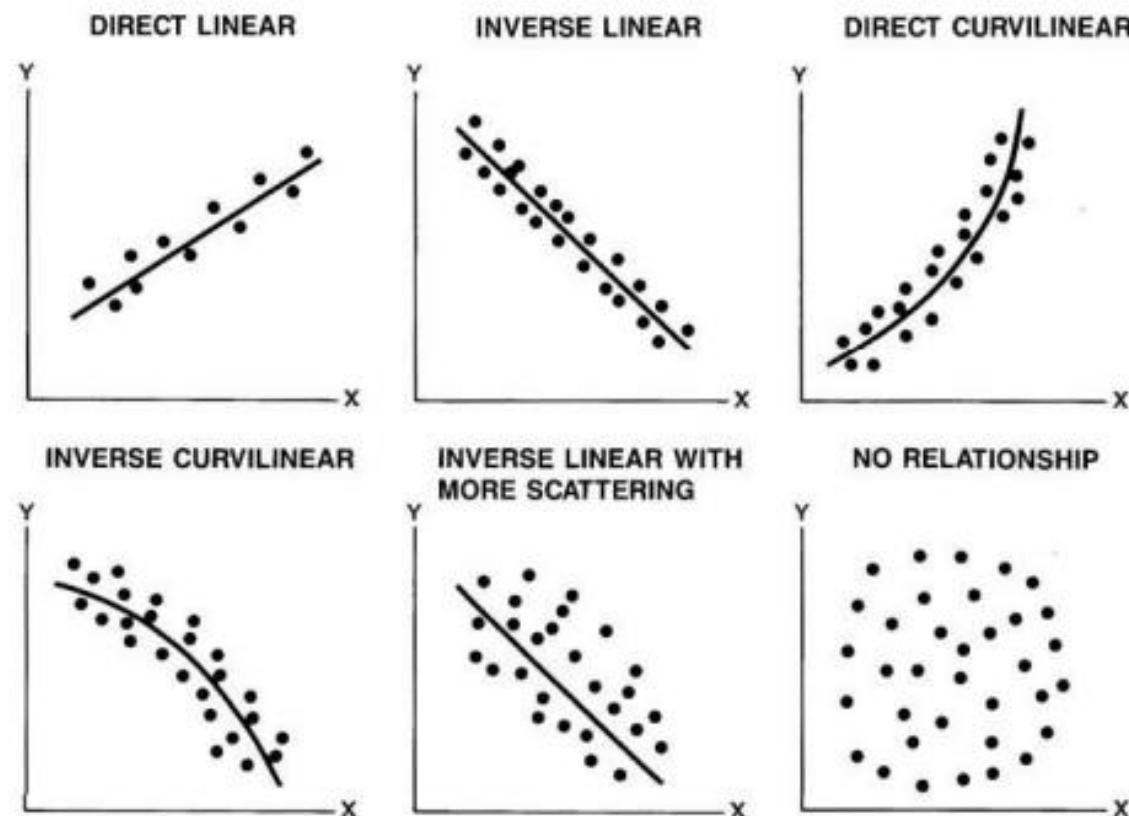
Correlation of RVs

- A measure which determines the standard change in one variable due to change in the other variable.
- Correlation is of two types, i.e. positive correlation or negative correlation.
- Correlation can take any value between -1 to +1, where in values close to +1 represents strong positive correlation and values close to -1 is an indicator of strong negative correlation.
- Measures of correlation:
 - Scatter diagram
 - Rank correlation coefficient

Correlation of RVs

- Correlation using scatter plot

Visual Relationship Between X and Y



Correlation of RVs

- Correlation coefficient

$$\text{Corr}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- Covariance, $\text{Cov}(X, Y)$ is dependent upon the units of X & Y.
- Correlation, $\text{Corr}(X, Y)$, scales covariance by the standard deviations of X & Y so that it lies between 1 & -1

$$\text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

Where σ is the Standard deviation

Common Distributions of RVs

- Uniform distribution
- Poisson distribution
- Normal distribution
- Standard normal distribution

Common Distributions of RVs

The Uniform Distribution

A random variable X is said to be *uniformly distributed* in $a \leq x \leq b$ if its density function is

$$f(x) = \begin{cases} 1/(b - a) & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

and the distribution is called a *uniform distribution*.

The distribution function is given by

$$F(x) = P(X \leq x) = \begin{cases} 0 & x < a \\ (x - a)/(b - a) & a \leq x < b \\ 1 & x \geq b \end{cases}$$

The mean and variance are, respectively,

$$\mu = \frac{1}{2}(a + b), \quad \sigma^2 = \frac{1}{12}(b - a)^2$$

Common Distributions of RVs

The Poisson Distribution

Let X be a discrete random variable that can take on the values $0, 1, 2, \dots$ such that the probability function of X is given by

$$f(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 0, 1, 2, \dots \quad (13)$$

where λ is a given positive constant. This distribution is called the *Poisson distribution* (after S. D. Poisson, who discovered it in the early part of the nineteenth century), and a random variable having this distribution is said to be *Poisson distributed*.

Mean	$\mu = \lambda$
Variance	$\sigma^2 = \lambda$
Standard deviation	$\sigma = \sqrt{\lambda}$

When p is small and n is fixed, Mean = $\lambda = np$, where

- n is the Number of Trials
- p is Probability of Success

Common Distributions of RVs

The Normal Distribution

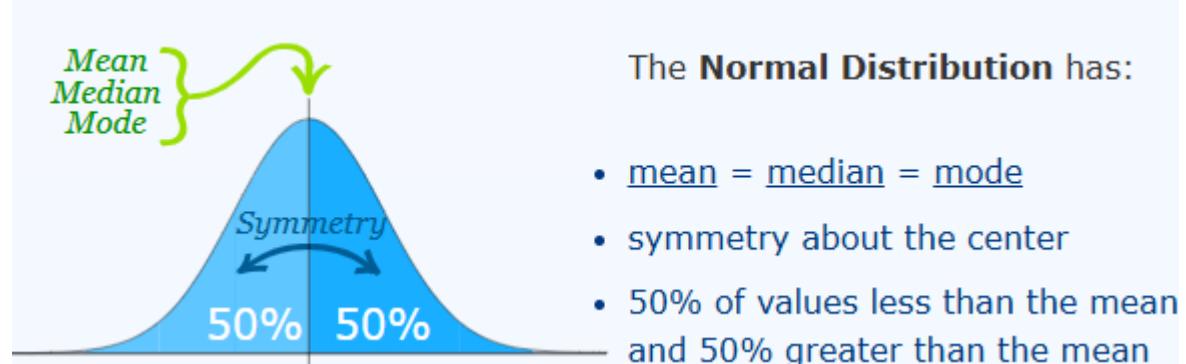
One of the most important examples of a continuous probability distribution is the *normal distribution*, sometimes called the *Gaussian distribution*. The density function for this distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty \quad (4)$$

where μ and σ are the mean and standard deviation, respectively. The corresponding distribution function is given by

$$F(x) = P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-(v-\mu)^2/2\sigma^2} dv \quad (5)$$

If X has the distribution function given by (5), we say that the random variable X is *normally distributed* with mean μ and variance σ^2 .



Common Distributions of RVs

Standard normal distribution, also known as the z-distribution

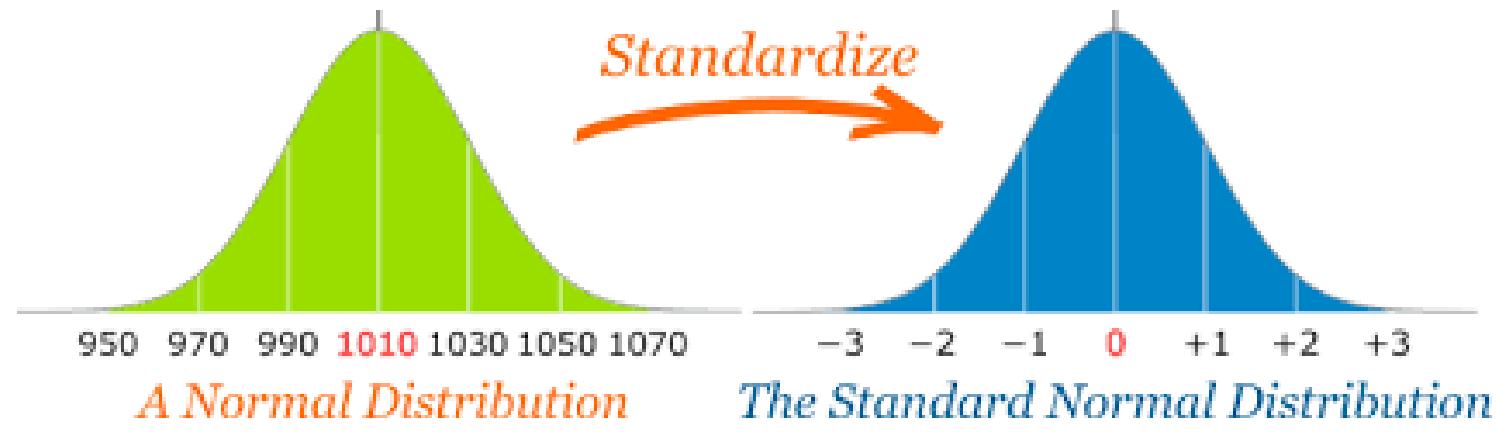
- In this distribution, the **mean (average)** is 0 and the **standard deviation (a measure of spread)** is 1.
- This creates a **bell-shaped curve** that is symmetrical around the mean ie. 0.
- The random variable of a standard normal distribution is known as the standard score or a z-score.

$$z = (X - \mu) / \sigma$$

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Where

$$-\infty < z < \infty$$



Central Limit Theorem

When large samples usually greater than thirty are taken into consideration then the distribution of sample arithmetic mean approaches the normal distribution irrespective of the fact that random variables were originally distributed normally or not.

Let us assume we have a random variable X.

Let σ be its standard deviation and μ is the mean of the random variable.

Now as per the Central Limit Theorem, the sample mean \bar{X} will approximate to the normal distribution which is given as $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$.

Central Limit Theorem Formula

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Sample Mean = Population Mean = μ

Sample Standard Deviation = $\frac{\text{Standard Deviation}}{n}$

OR

Sample Standard Deviation = $\frac{\sigma}{\sqrt{n}}$

Statistics

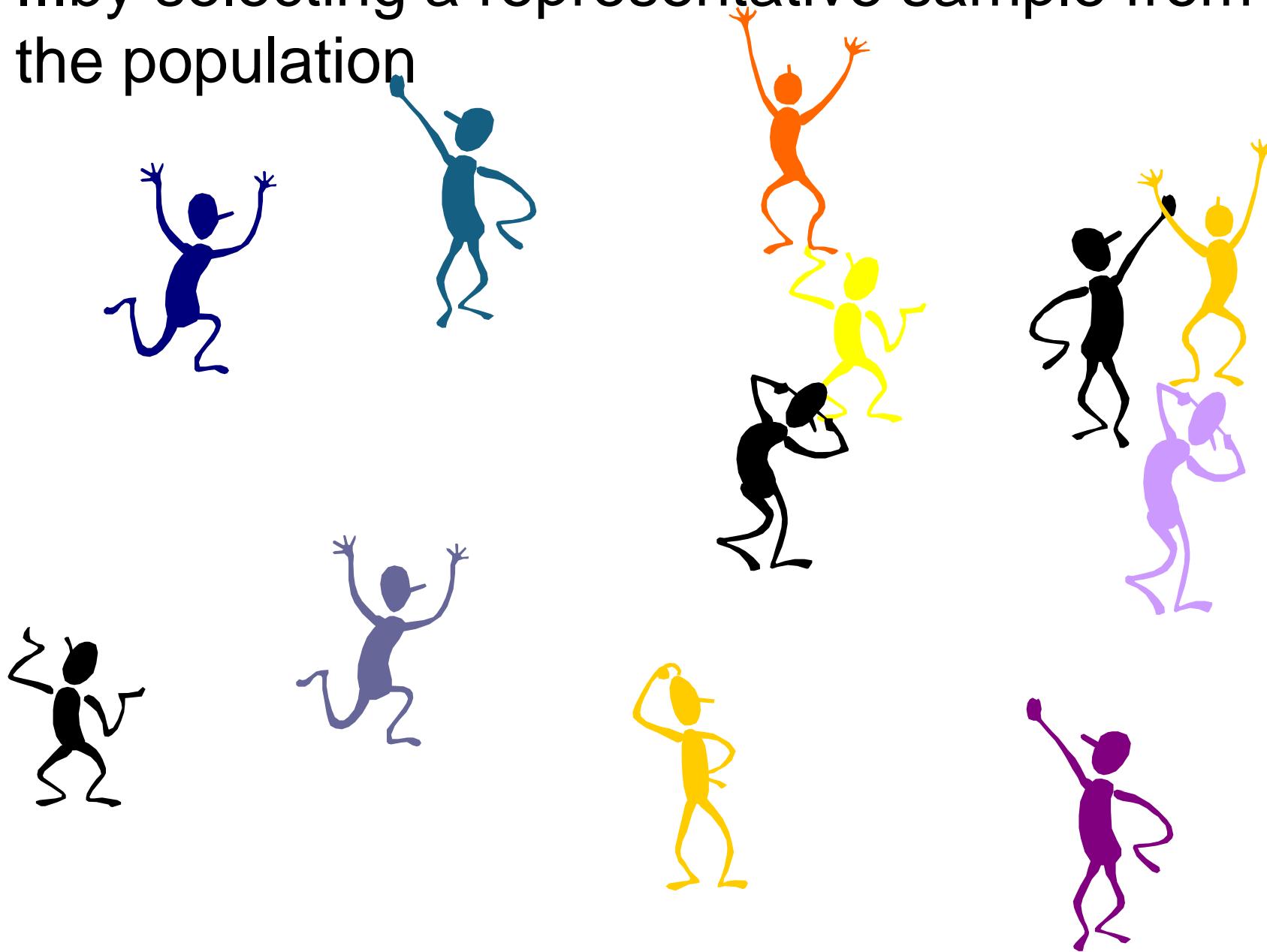
Sampling

- **Population** – A group that includes all the cases (individuals, objects, or groups) in which the researcher is interested.
- **Sample** – A relatively small subset from a population.
- **Simple Random Sample** – A sample designed in such a way as to ensure that
 - (1) every member of the population has an equal chance of being chosen and
 - (2) every combination of N members has an equal chance of being chosen.
- This can be done using a computer, calculator, or a table of random numbers

Population inferences can be made...

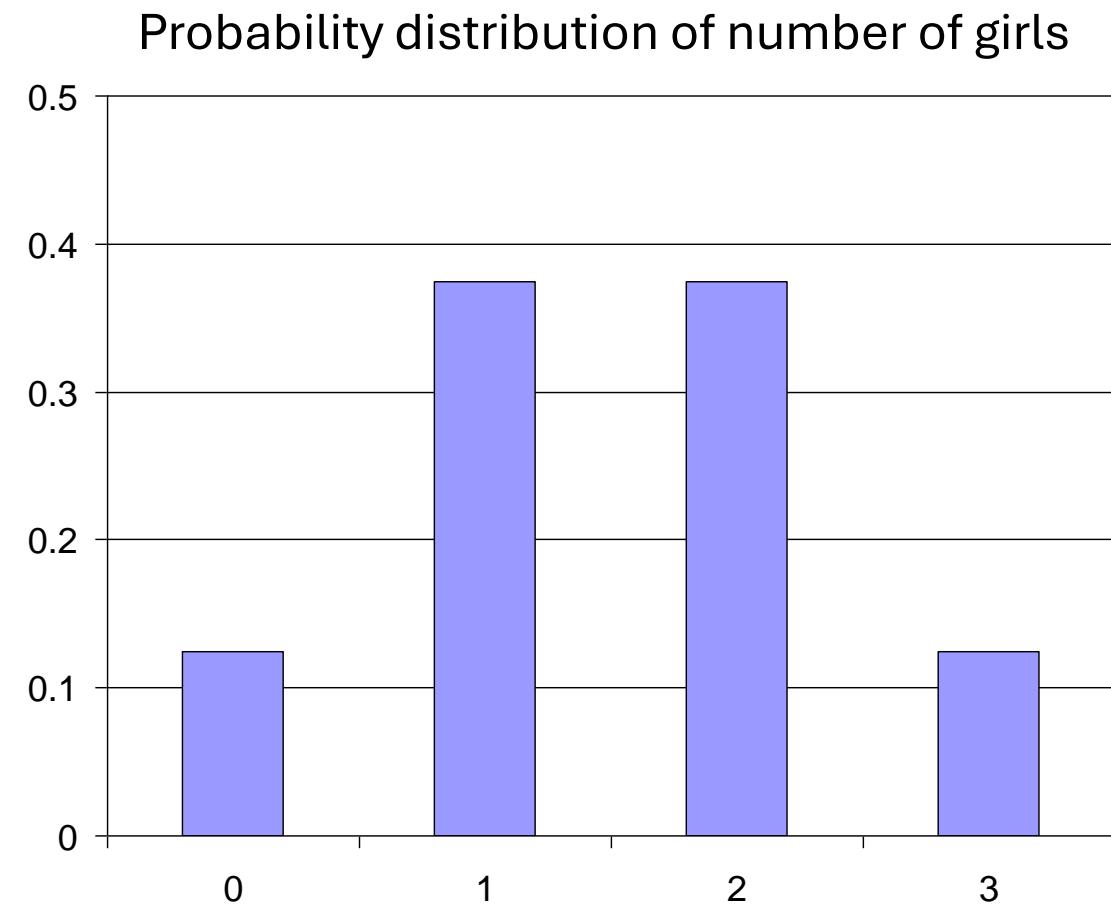


...by selecting a representative sample from
the population



How about family of three?

Num Girls	child #1	child #2	child #3
0	B	B	B
1	B	B	G
1	B	G	B
1	G	B	B
2	B	G	G
2	G	B	G
2	G	G	B
3	G	G	G



Probability distributions: Permutations

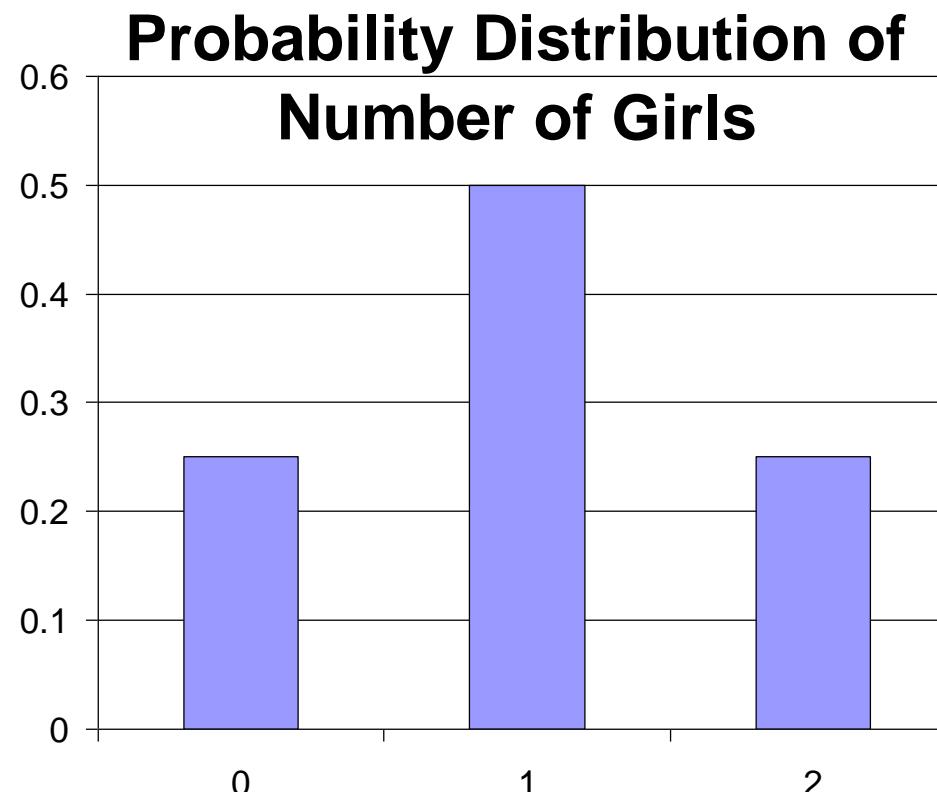
What is the probability distribution of number of girls in families with two children?

2 GG

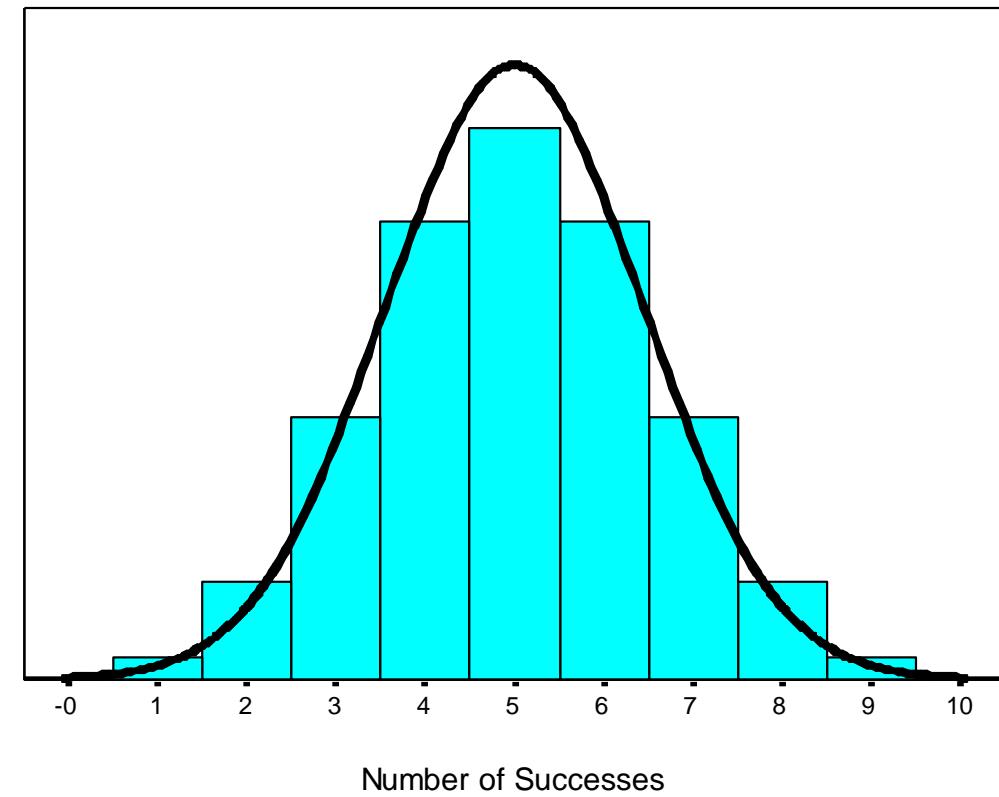
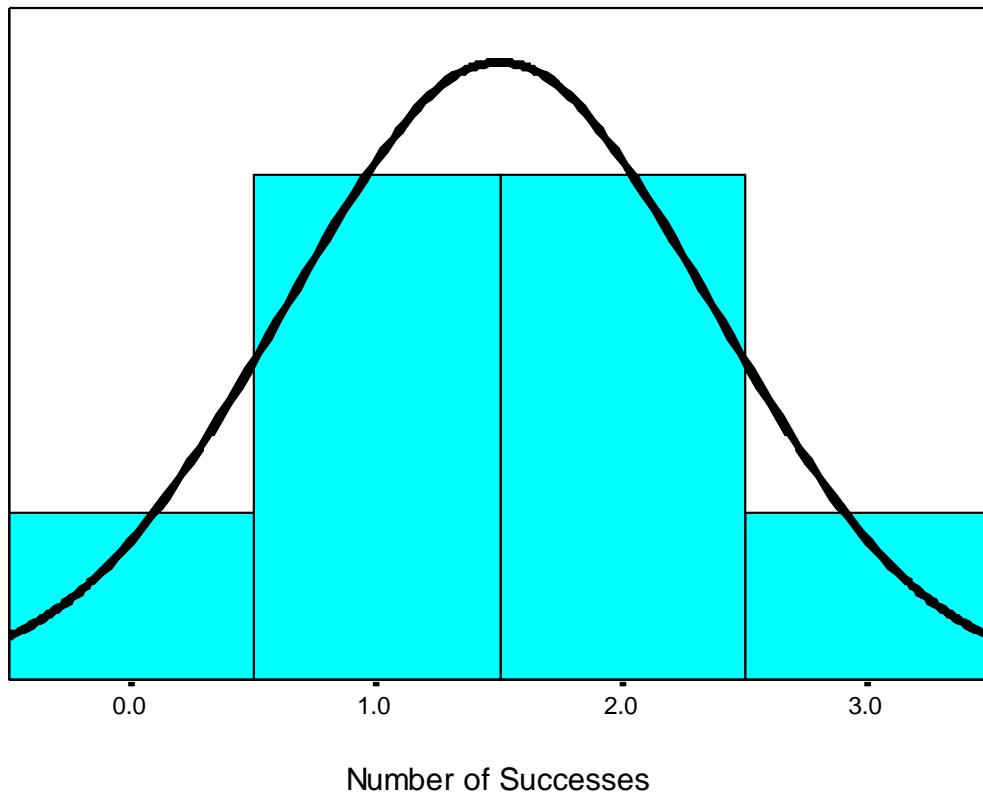
1 BG

1 GB

0 BB



As family size increases, the distribution looks more and more normal.



Coin toss

- Toss a coin 30 times
- Tabulate results
- Think of the coin tosses as samples of all possible coin tosses

Sampling Distribution

- Imagine repeatedly taking samples of the same size from the large population and calculating a statistic (like the mean or variance) for each sample.
- The probability distribution of these calculated statistics is called the sampling distribution.
- Aim of sampling
 - Reduces cost of research (e.g. political polls)
 - Generalize about a larger population (e.g., benefits of sampling city r/t neighborhood)
 - In some cases (e.g. industrial production) analysis may be destructive, so sampling is needed

Central Limit Theorem

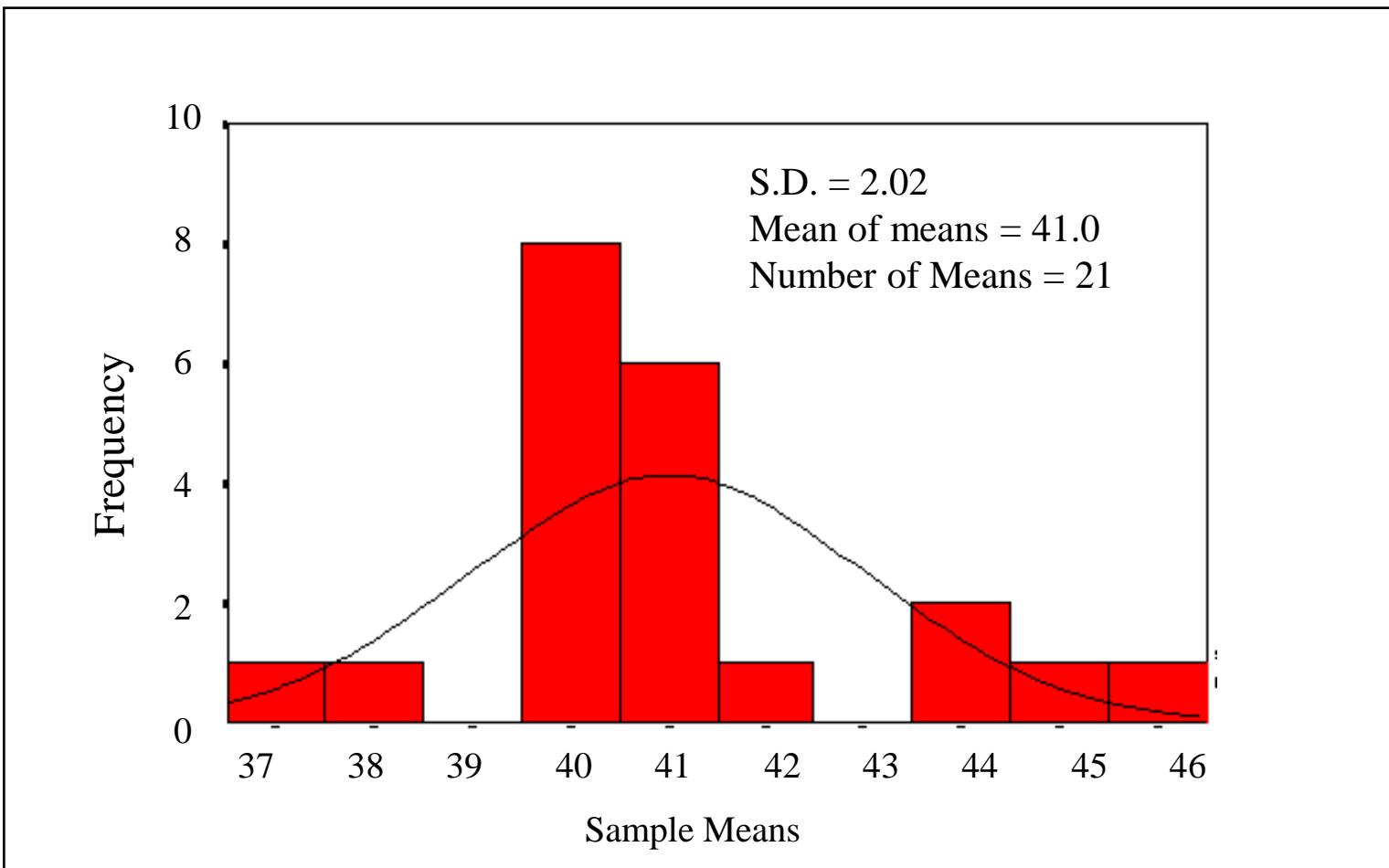
- No matter what we are measuring, the distribution of any measure across all possible samples we could take, approximates a normal distribution, as long as the number of cases in each sample is about 30 or larger.

If we repeatedly drew samples from a population and calculated the mean of a variable or a percentage, those sample means or percentages would be normally distributed.

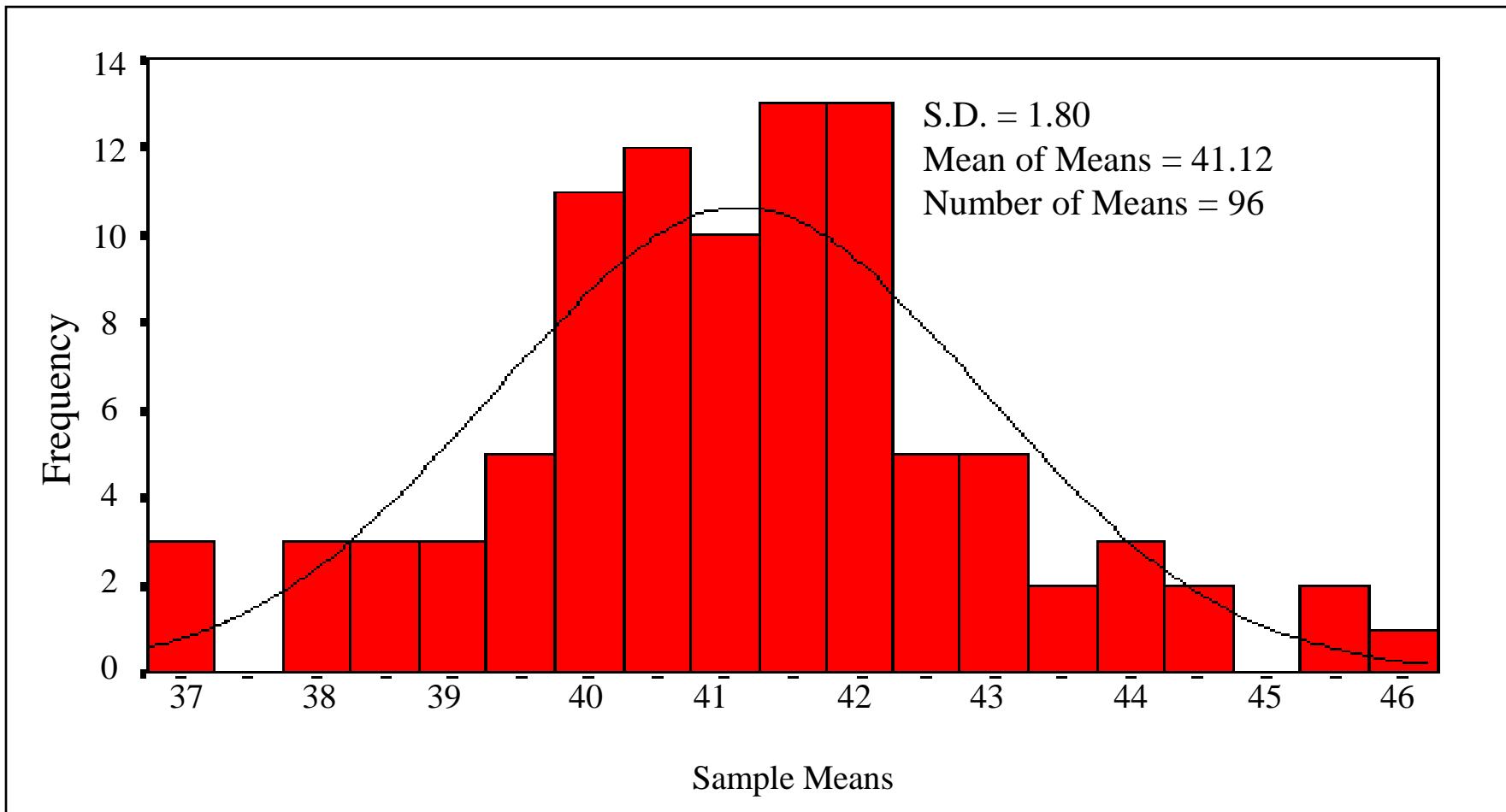
The Mean and Standard Deviation of the Sample Mean

- Suppose we wish to estimate the mean μ of a population. In actual practice we would typically take just one sample.
- Imagine however that we take sample after sample, all with same sample size n , and compute the sample mean \bar{x} each time.
- The sample mean \bar{x} is a random variable: it varies from sample to sample in a way that cannot be predicted with certainty.
- Consider \bar{X} , as a random variable of the sample mean, and write x for the values that it takes.
- The random variable \bar{X} has a mean, denoted $\mu_{\bar{x}}$, and a standard deviation, denoted $\sigma_{\bar{x}}$.

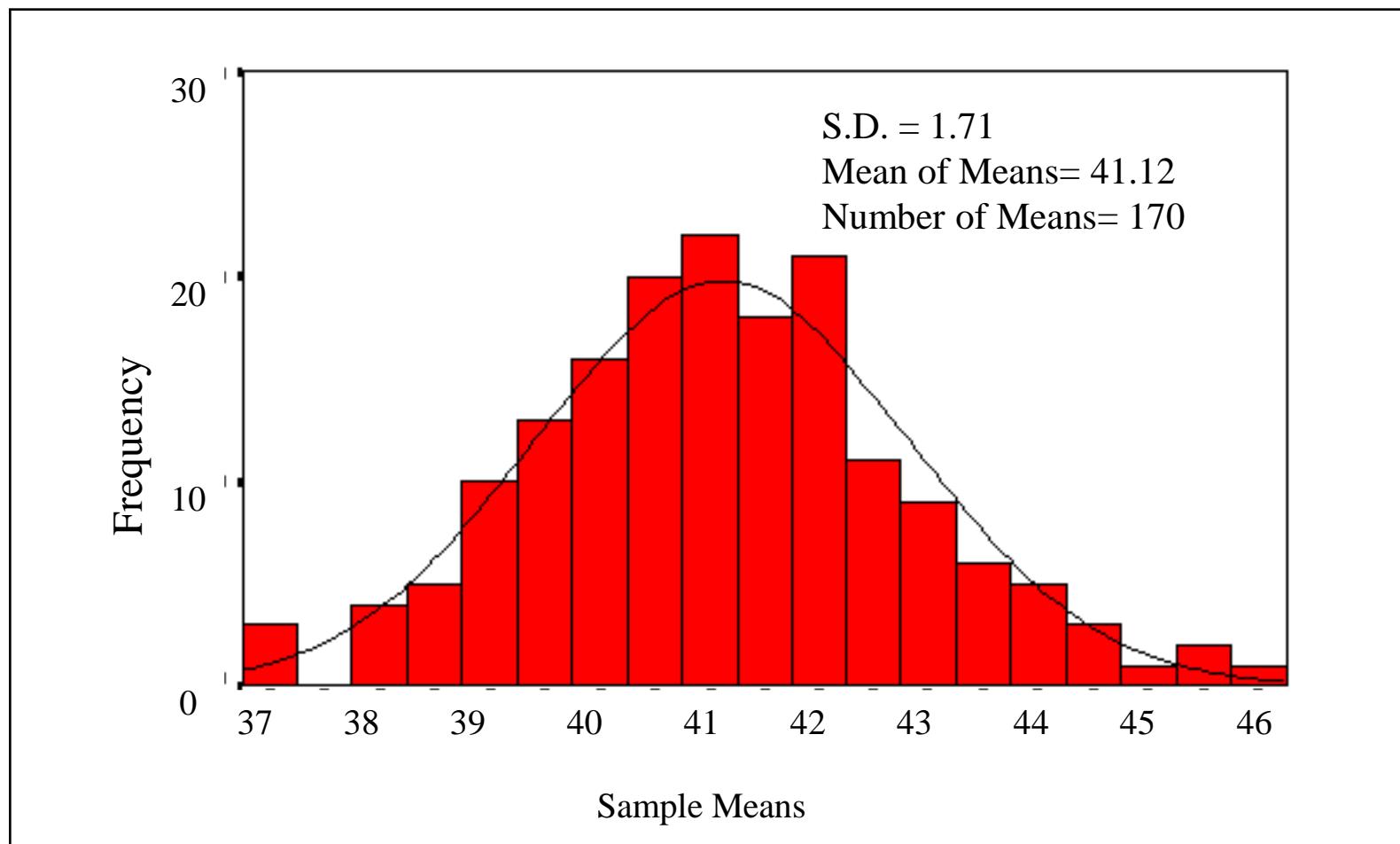
Distribution of Sample Means with 21 Samples



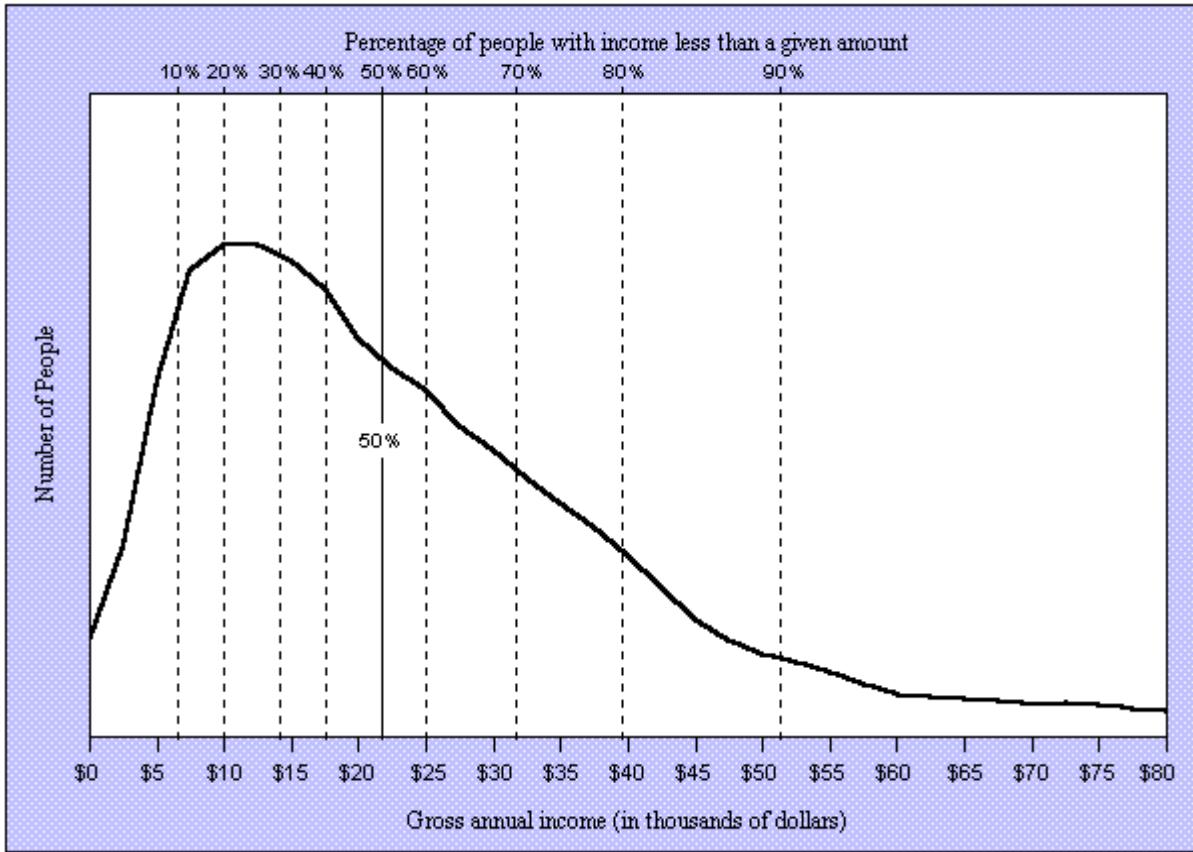
Distribution of Sample Means with 96 Samples



Distribution of Sample Means with 170 Samples

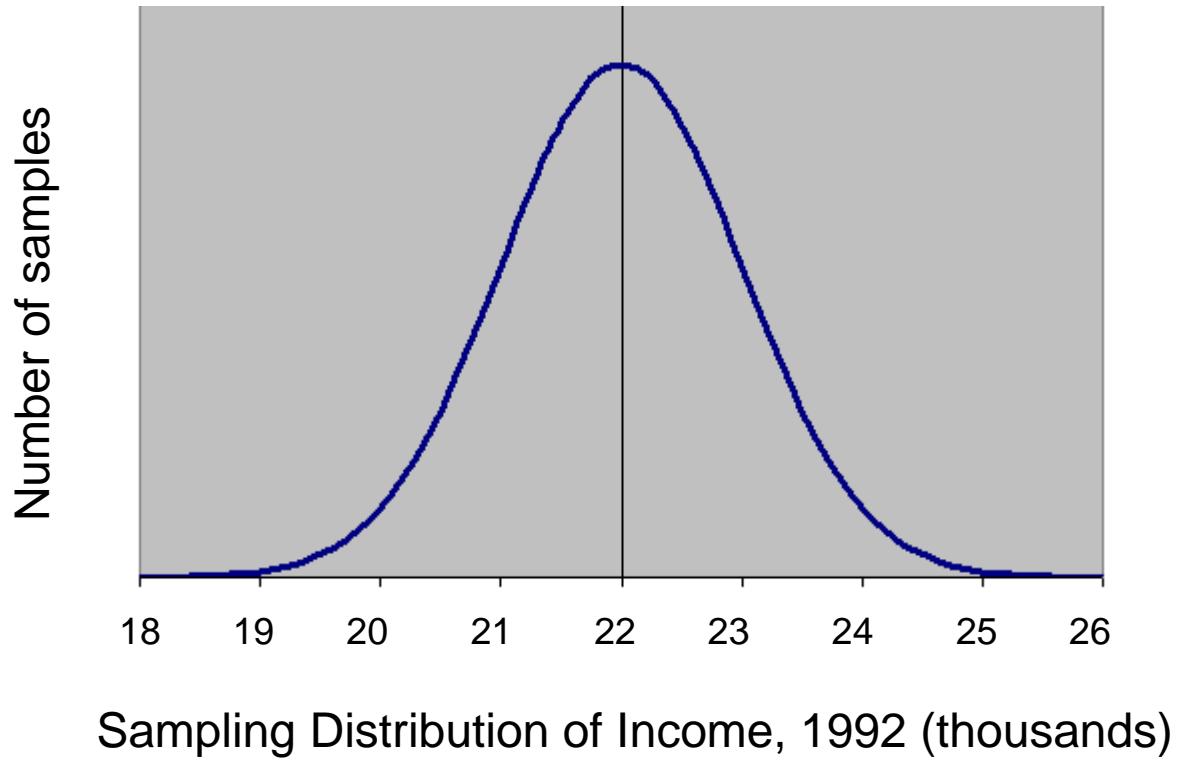


Most empirical distributions are not normal:



U.S. Income distribution 1992

But the sampling distribution of mean income over many samples *is* normal



Central Limit Theorem

When large samples usually greater than thirty are taken into consideration then the distribution of sample arithmetic mean approaches the normal distribution irrespective of the fact that random variables were originally distributed normally or not.

Let us assume we have a random variable X.

Let σ be its standard deviation and μ is the mean of the random variable.

Now as per the Central Limit Theorem, the sample mean \bar{X} will approximate to the normal distribution which is given as $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$.

Central Limit Theorem Formula

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

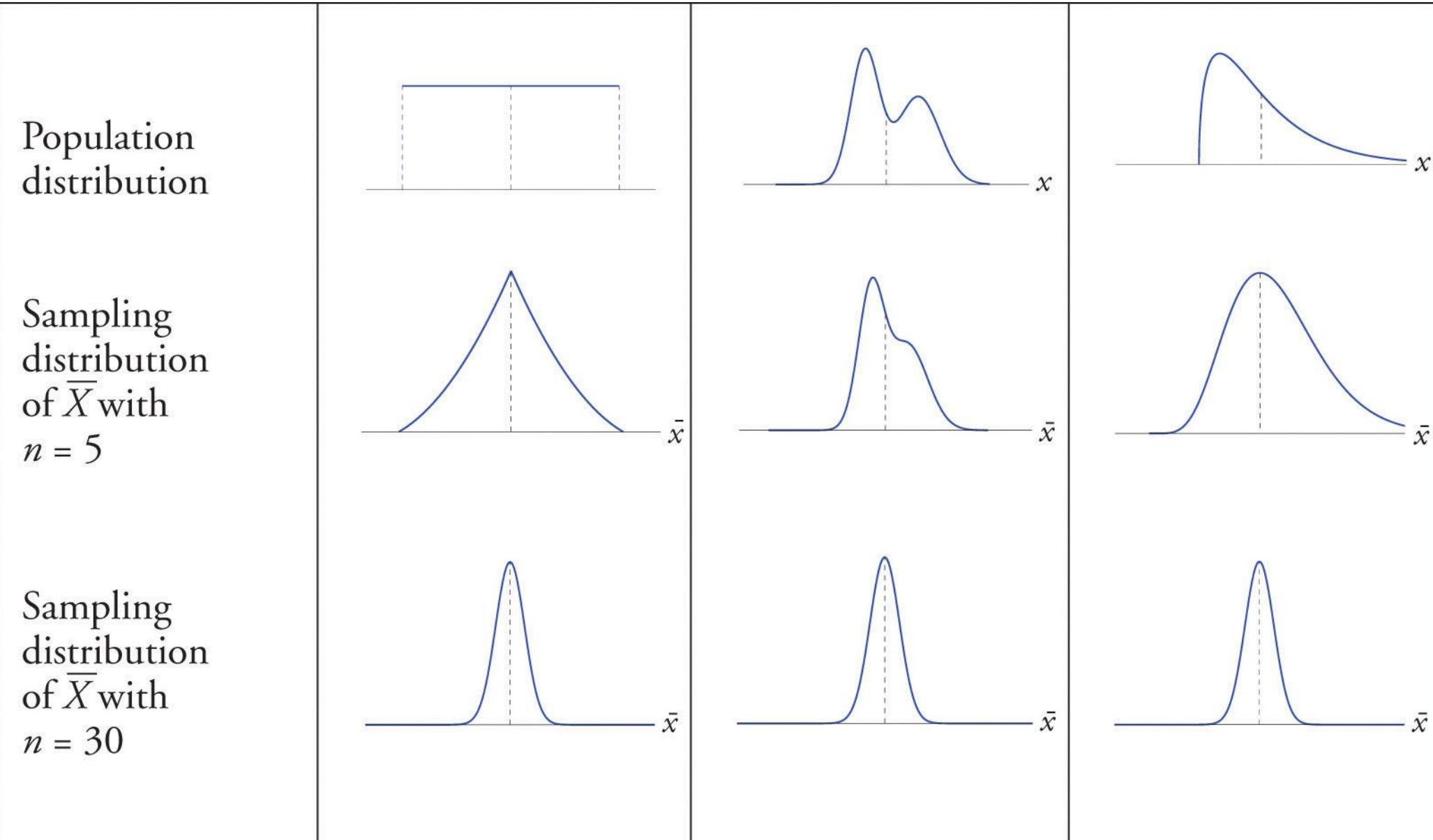
Sample Mean = Population Mean = μ

Sample Standard Deviation = $\frac{\text{Standard Deviation}}{n}$

OR

Sample Standard Deviation = $\frac{\sigma}{\sqrt{n}}$

Central Limit Theorem



Q1.

The mean and standard deviation of the tax value of all vehicles registered in a certain state are $\mu = \$13,525$ and $\sigma = \$4,180$. Suppose random samples of size 100 are drawn from the population of vehicles. What are the mean $\mu_{\bar{X}}$ and standard deviation $\sigma_{\bar{X}}$ of the sample mean \bar{X} ?

Solution

Since $n = 100$, the formulas yield

$$\mu_{\bar{X}} = \mu = \$13,525$$

and

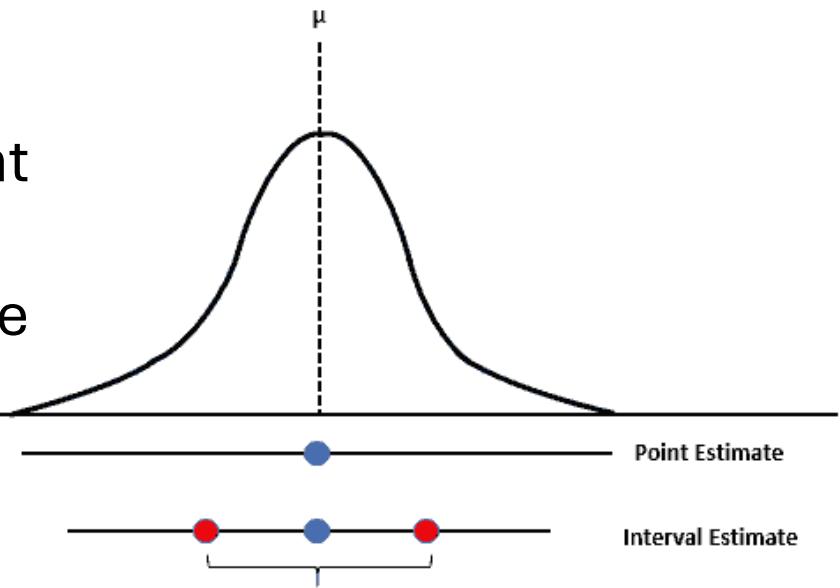
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$4,180}{\sqrt{100}} = \$418$$

Point Estimate and Interval Estimate

- A **point estimate** is a single value estimate of a parameter. For instance, a sample mean is a point estimate of a population mean.
- A point estimate is a sample statistic calculated using the sample data to estimate the most likely value of the corresponding unknown population parameter. In other words, we derive the point estimate from a single value in the sample and use it to estimate the population value.
- Take a sample, find \bar{x} . It is a close approximation of μ . But, depending on your sample size, that may not be a good point estimate.
- In fact, the probability that a single sample statistic is equal to the population parameter is very unlikely.

Point Estimate and Interval Estimate

- An interval estimate gives you a range of values where the parameter is expected to lie.
- A confidence interval estimate is a range of values constructed from sample data so that the population parameter will likely occur within the range at a specified probability. Accordingly, the specified probability is the level of confidence.
- Broader and probably more accurate than a point estimate
- Any parameter estimate that is based on a sample statistic has some amount of sampling error.



Point Estimate and Interval Estimate

A Confidence interval is used to express the precision and ambiguity of a particular sampling method.

- A confidence *interval* is a range of values that probably contain the population mean.
- A Confidence level is a percentage of certainty that, in any given sample, that confidence interval will contain the population means.
- The Point estimate is a statistic (value from a sample) used to estimate a parameter (value from the population).
- The margin of error is the maximum expected difference between the actual population parameter and a sample estimate of the parameter. In other words, it is the range of values above and below sample statistics.

$$\mu = \bar{x} \pm z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

Diagram illustrating the components of a confidence interval formula:

- Point Estimate**: \bar{x}
- Confidence Level**: $z_{\alpha/2}$
- Margin of Error**: $\frac{\sigma}{\sqrt{n}}$

Arrows point from each label to its corresponding term in the formula.

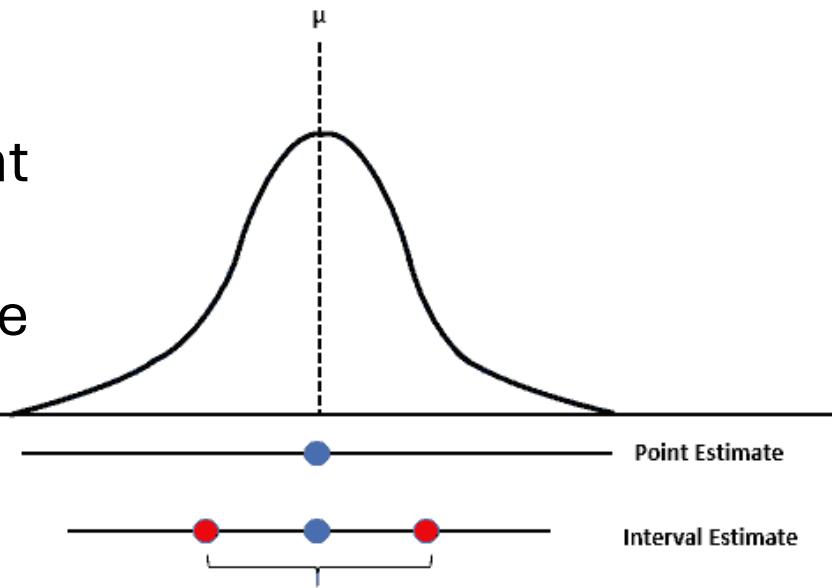
Statistics

Point Estimate and Interval Estimate

- A **point estimate** is a single value estimate of a parameter. For instance, a sample mean is a point estimate of a population mean.
- A point estimate is a sample statistic calculated using the sample data to estimate the most likely value of the corresponding unknown population parameter. In other words, we derive the point estimate from a single value in the sample and use it to estimate the population value.
- Take a sample, find \bar{x} . It is a close approximation of μ . But, depending on your sample size, that may not be a good point estimate.
- In fact, the probability that a single sample statistic is equal to the population parameter is very unlikely.

Point Estimate and Interval Estimate

- An interval estimate gives you a range of values where the parameter is expected to lie.
- A confidence interval estimate is a range of values constructed from sample data so that the population parameter will likely occur within the range at a specified probability. Accordingly, the specified probability is the level of confidence.
- Broader and probably more accurate than a point estimate
- Any parameter estimate that is based on a sample statistic has some amount of sampling error.



Point Estimate and Interval Estimate

A Confidence interval is used to express the precision and ambiguity of a particular sampling method.

- A confidence *interval* is a range of values that probably contain the population mean.
- A Confidence level is a percentage of certainty that, in any given sample, that confidence interval will contain the population means.
- The Point estimate is a statistic (value from a sample) used to estimate a parameter (value from the population).
- The margin of error is the maximum expected difference between the actual population parameter and a sample estimate of the parameter. In other words, it is the range of values above and below sample statistics.

$$\mu = \bar{x} \pm z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

Diagram illustrating the components of a confidence interval formula:

- Point Estimate**: \bar{x}
- Confidence Level**: $z_{\alpha/2}$
- Margin of Error**: $\frac{\sigma}{\sqrt{n}}$

The formula combines the point estimate with the margin of error to form the confidence interval.

Point Estimation

- Here, we assume that θ is an unknown parameter to be estimated.
- For example, θ might be the expected value of a random variable, $\theta = EX$. Θ is a fixed (non-random) quantity.
- To estimate θ , we define a point estimator $\hat{\Theta}$ that is a function of the random sample, i.e.,

$$\hat{\Theta} = h(X_1, X_2, \dots, X_n).$$

For example, if $\theta = EX$, we may choose $\hat{\Theta}$ to be the sample mean

$$\hat{\Theta} = \bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

Point Estimation

Mean (\bar{x}) → Estimates Population Mean (μ)

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$$

Variance (s^2) → Estimates Population Variance (σ^2)

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Standard Deviation (s) → Estimates Population Standard Deviation (σ)

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

Proportion (\hat{p}) → Estimates Population Proportion (p)

$$\hat{p} = \frac{x}{n}$$

Properties of Estimators

- Estimators should be **unbiased**.
 - expected value equals the true parameter value.
- The estimator should be **efficient**.
 - it has the **lowest variance** among all unbiased estimators of a parameter.
- An estimator should be **consistent**.
 - as the sample size increases, the estimated value gets closer to the true population parameter.
 - More data improves the accuracy of estimation.

Evaluating Estimators

- Three main desirable properties for point estimators
 1. The **bias** of an estimator $\hat{\Theta}$ tells us on average how far $\hat{\Theta}$ is from the real value of θ .
- An estimator $\hat{\theta}$ is **unbiased** if its expected value is equal to the true population parameter (θ):
$$E(\hat{\theta}) = \theta$$
- Example: The sample mean \bar{x} is an unbiased estimator of the population mean μ :
$$E(\bar{X}) = \mu$$

ii. *consistency*

- An estimator $\hat{\theta}$ is **consistent** if it gets **closer to the true parameter (θ) as the sample size (n) increases.**

Let $\hat{\Theta}_1, \hat{\Theta}_2, \dots, \hat{\Theta}_n, \dots$, be a sequence of point estimators of θ . We say that $\hat{\Theta}_n$ is a **consistent** estimator of θ , if

$$\lim_{n \rightarrow \infty} P(|\hat{\Theta}_n - \theta| \geq \epsilon) = 0, \text{ for all } \epsilon > 0.$$

- Example: The **sample mean (\bar{x})** is a consistent estimator of μ because as we take larger samples, it converges to μ .

iii. Efficiency

- Among multiple unbiased estimators, the **most efficient** estimator has the **smallest variance**.

$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2) \Rightarrow \hat{\theta}_1 \text{ is more efficient}$$

Example: If we have two estimators of μ , the one with the smaller variance is preferred.

Estimator 1: $Var(\hat{\theta}_1) = 5$

Estimator 2: $Var(\hat{\theta}_2) = 2$

Hypothesis Testing

It refers to

- Making an assumption, called hypothesis, about a population parameter.
- Collecting sample data and calculating sample statistic.
- Using the sample statistic to evaluate the hypothesis (how likely is it that our hypothesized parameter is correct).
- To test the validity of our assumption we determine the difference between the hypothesized parameter value and the sample value.

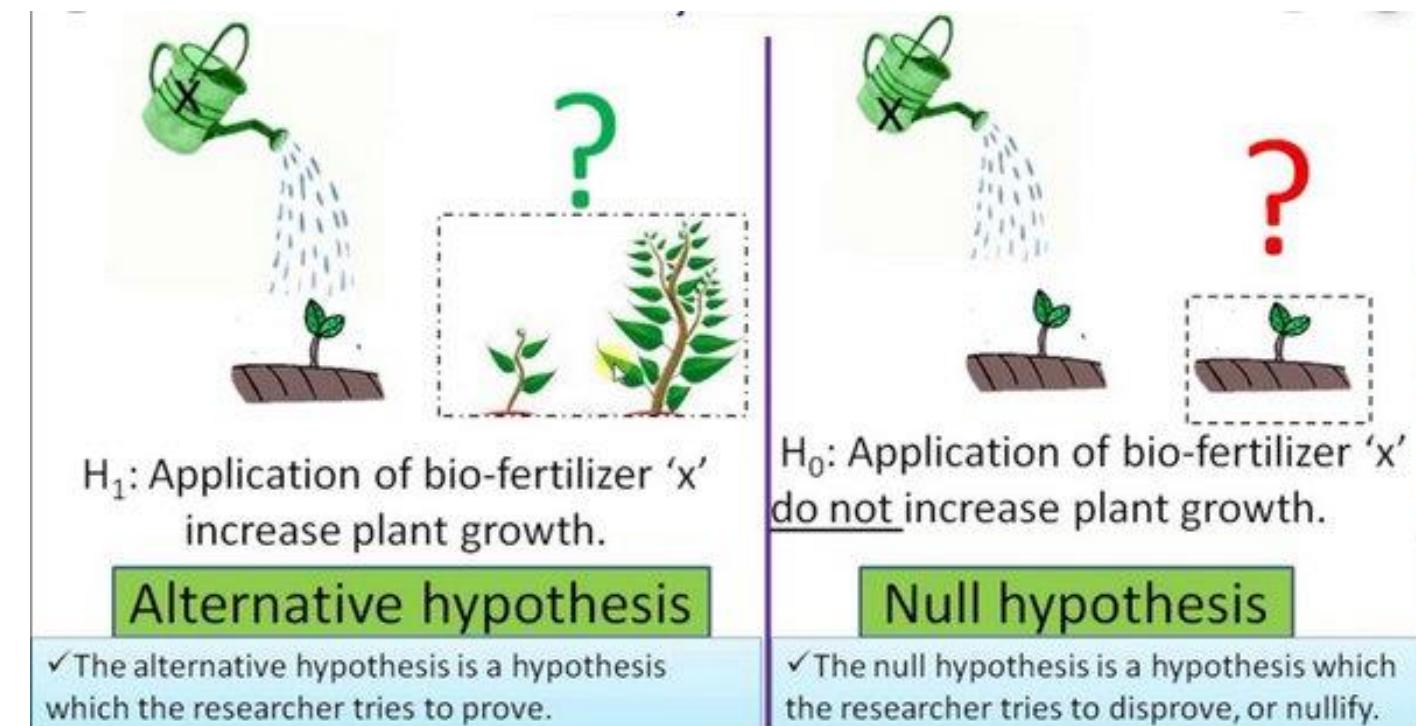
Hypothesis Testing

Example:

- A pharmaceutical company might be interested in knowing if a new drug is effective in treating a disease. Here, there are two hypotheses.
- The first one is that the drug is not effective (hypotheses H₀), while the second hypothesis is that the drug is effective (hypotheses H₁).
- The hypothesis H₀ is called the **null hypothesis** and the hypothesis H₁ is called the **alternative hypothesis**.

Hypothesis Testing

The null hypothesis represents the default assumption that no significant difference or relationship exists between the studied variables. In contrast, the alternative hypothesis represents the claim or hypothesis the researcher is testing.



Hypothesis Testing

You have a coin and you would like to check whether it is fair or not. More specifically, let θ be the probability of heads, $\theta = P(H)$. You have two hypotheses:

H_0 (the null hypothesis): The coin is fair, i.e. $\theta = \theta_0 = \frac{1}{2}$.

H_1 (the alternative hypothesis): The coin is not fair, i.e., $\theta \neq \frac{1}{2}$.

We need to design a test to either accept H_0 or H_1 . To check whether the coin is fair or not, we perform the following experiment. We toss the coin 100 times and record the number of heads. Let X be the number of heads that we observe, so

$$X \sim \text{Binomial}(100, \theta).$$

Now, if H_0 is true, then $\theta = \theta_0 = \frac{1}{2}$, so we expect the number of heads to be close to 50. Thus, intuitively we can say that if we observe close to 50 heads we should accept H_0 , otherwise we should reject it. More specifically, we suggest the following criteria: If $|X - 50|$ is less than or equal to some threshold, we accept H_0 . On the other hand, if $|X - 50|$ is larger than the threshold we reject H_0 and accept H_1 . Let's call that threshold t .

If $|X - 50| \leq t$, accept H_0 .

If $|X - 50| > t$, accept H_1 .

Hypothesis Testing

- **Level of significance:** It refers to the degree of significance in which we accept or reject the null hypothesis. 100% accuracy is not possible for accepting a hypothesis, so we select a level of significance. This is normally denoted with α (alpha) and generally, it is 0.05 or 5% which means your output should be 95% confident to give a similar kind of result in each sample.
- **Test Statistic:** Test statistic is the number that helps you decide whether your result is significant. It's calculated from the sample data you collect it could be used to test if a machine learning model performs better than a random guess.
- **Critical value:** Critical value is a boundary or threshold that helps you decide if your test statistic is enough to reject the null hypothesis

Hypothesis Testing

- **P-value:** The p-value is the probability of observing a test statistic given that the null hypothesis is true.
 - A **small p-value** usually less than 0.05 means the results are unlikely to be due to random chance so we reject the null hypothesis.
 - A **large p-value** means the results could easily happen by chance so we don't reject the null hypothesis.

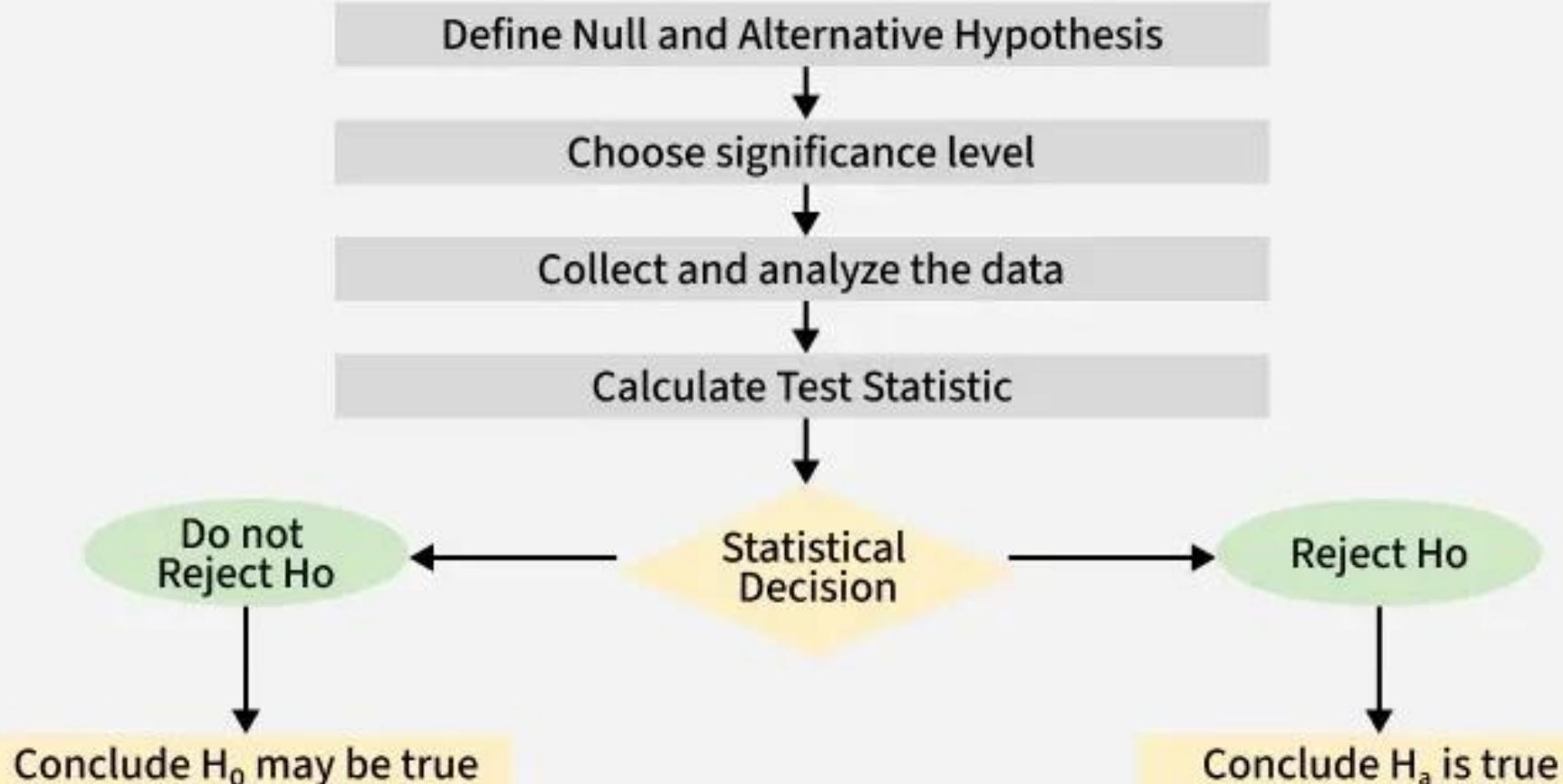
Hypothesis Testing

In hypothesis testing Type I and Type II errors are two possible errors that can happen when we are finding conclusions about a population based on a sample of data. These errors are associated with the decisions we made regarding the null hypothesis and the alternative hypothesis.

- **Type I error:** When we reject the null hypothesis although that hypothesis was true. Type I error is denoted by alpha(α).
- **Type II errors:** When we accept the null hypothesis, but it is false. Type II errors are denoted by beta(β).

	Null Hypothesis is True	Null Hypothesis is False
Null Hypothesis is True (Accept)	Correct Decision	Type II Error (False Negative)
Alternative Hypothesis is True (Reject)	Type I Error (False Positive)	Correct Decision

Hypothesis Testing



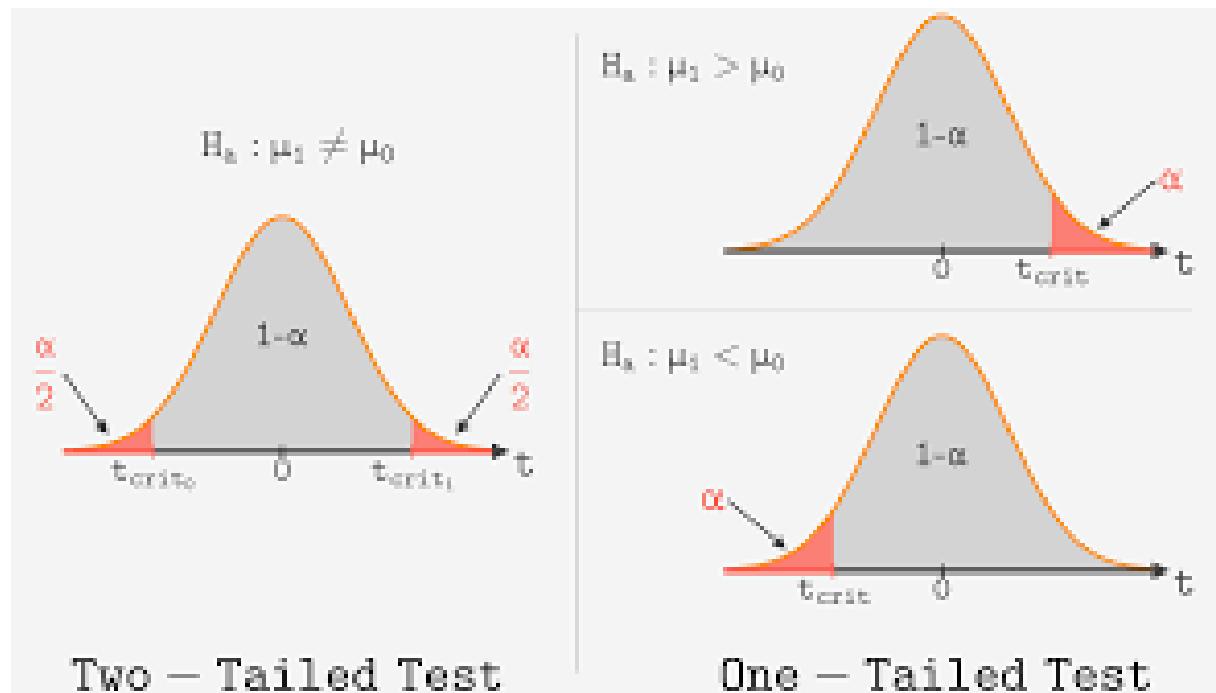
One-Sided (One-Tailed) Tests

- **Greater Than Test (Right-Tailed Test):** This type of one-sided test is used when you want to determine if a parameter or effect is greater than a specified value.
 - **Null Hypothesis (H_0):** The parameter is less than or equal to the specified value.
 - **Alternative Hypothesis (H_1 or H_a):** The parameter is greater than the specified value.
 - **Example:** Testing if a new drug improves patient recovery time H_0 : The drug does not improve recovery time. H_a : The drug improves recovery time.
- **Less Than Test (Left-Tailed Test):** This one-sided test is used when you want to determine if a parameter or effect is less than a specified value.
 - **Null Hypothesis (H_0):** The parameter is greater than or equal to the specified value.
 - **Alternative Hypothesis (H_1 or H_a):** The parameter is less than the specified value.
 - **Example:** Testing if a manufacturing process meets quality standards. H_0 : The process meets quality standards. H_a : The process does not meet quality standards.

Two-Sided (Two-Tailed) Tests

Two-Sided Test: This type of test is used when you want to determine if a parameter or effect is significantly different from a specified value, without specifying whether it's greater or less than that value.

- **Null Hypothesis (H_0):** The parameter is equal to the specified value.
- **Alternative Hypothesis (H_1 or H_a):** The parameter is not equal to the specified value.
- **Example:** Testing if a coin is fair (i.e., equally likely to land heads or tails). H_0 : The coin is fair. H_a : The coin is not fair.



Statistics

z-Statistics

It is used when population means and standard deviations are known. The formula of z-statistics is given by:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where

- \bar{x} is the sample mean,
- μ represents the population mean,
- σ is the standard deviation
- and n is the size of the sample.

z-Statistics

Q1. A company manufactures light bulbs and claims that the average lifespan of their bulbs is 1000 hours. A consumer group wants to test this claim. They randomly sample 64 bulbs and find that the sample mean lifespan is 980 hours. Assume the population standard deviation is known to be 100 hours.

1. State the hypotheses:

- Null hypothesis (H_0): The average lifespan of the bulbs is 1000 hours. ($\mu = 1000$)
- Alternative hypothesis (H_a): The average lifespan of the bulbs is not 1000 hours. ($\mu \neq 1000$)

2. Determine the level of significance:

- Let's choose a significance level of 0.05 (alpha = 0.05). This means we are willing to accept a 5% chance of rejecting the null hypothesis when it is actually true.

3. Calculate the test statistic:

- The formula for the z-test statistic is: $z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$
 - Where:
 - \bar{x} is the sample mean (980 hours)
 - μ is the population mean under the null hypothesis (1000 hours)
 - σ is the population standard deviation (100 hours)
 - n is the sample size (64)
- Plugging in the values: $z = (980 - 1000) / (100 / \sqrt{64}) = -1.6$

4. Find the critical value:

- This is a two-tailed test ($H_a: \mu \neq 1000$)
- For a two-tailed test with alpha = 0.05, the critical values are approximately ± 1.96 .

5. Make a decision:

- Compare test statistic (-1.6) to the critical values (± 1.96).
- Since the absolute value of the test statistic (1.6) is less than the critical value (1.96), we fail to reject the null hypothesis.

Conclusion:

At a 0.05 significance level, there is not enough evidence to conclude that the average lifespan of the light bulbs is different from 1000 hours.

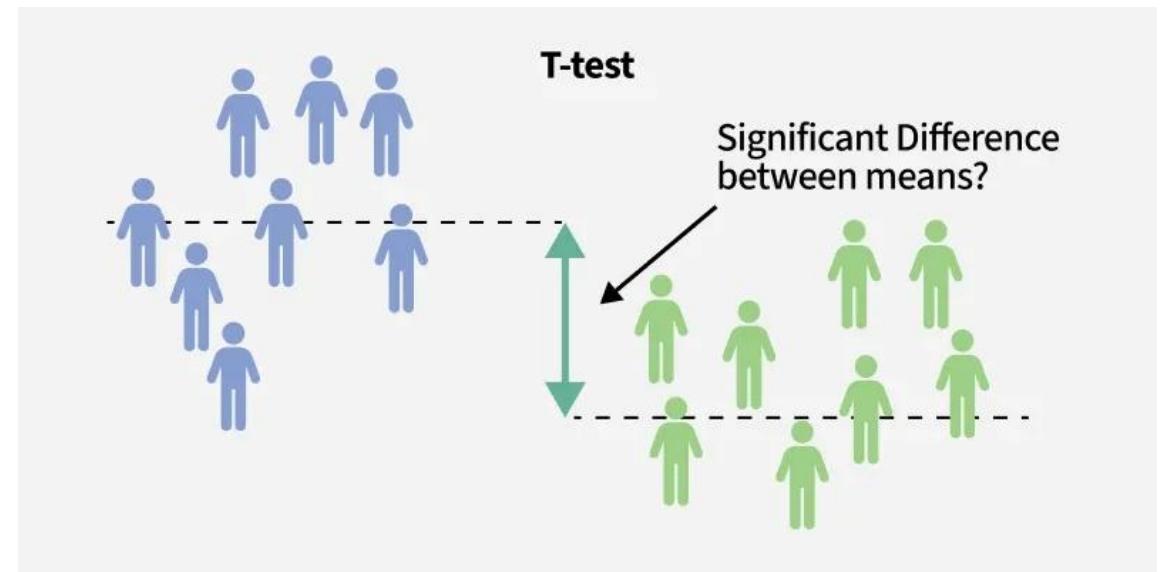
T-Statistics

T-test is used to compare the means of two datasets (e.g., experimental vs. control groups) to assess if the difference is statistically significant. It is used when $n < 30$ t-statistic calculation is given by:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

where

- t = t-score,
- \bar{x} = sample mean
- μ = population mean,
- s = standard deviation of the sample,
- n = sample size



T-Statistics

Suppose You want to compare the test scores of two groups of students:

- Group 1: 30 students who studied with Method A.
- Group 2: 30 students who studied with Method B.

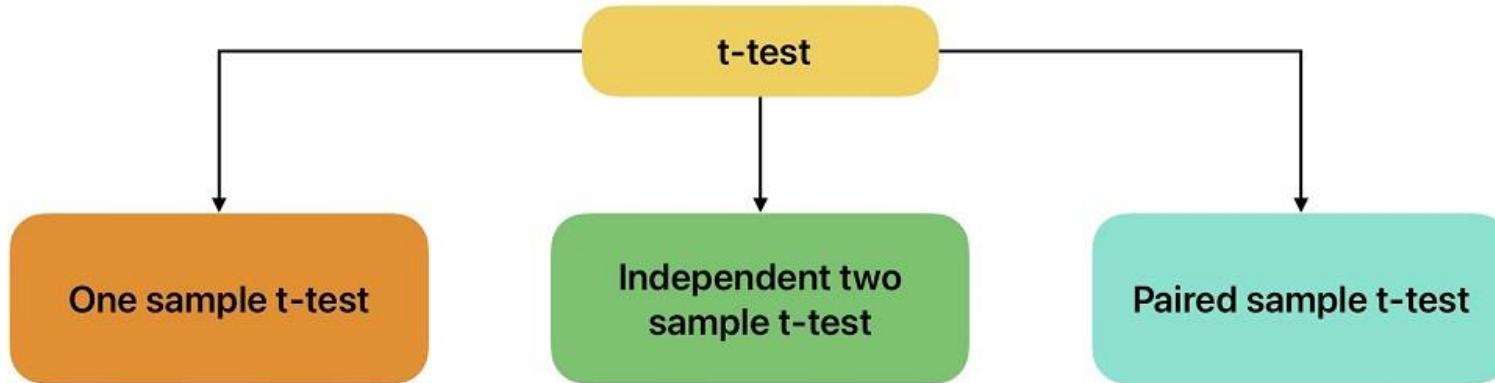
You use a **t-test** to check if there is a significant difference in the average test scores between the two.

The t-test is part of **hypothesis testing** where you start with an assumption the null hypothesis that the two-group means are the same. Then the test helps you decide if there's enough evidence to reject that assumption and conclude that the groups are different.

T-Statistics

- **Degree of freedom (df):** The degree of freedom tells us the number of independent variables used for calculating the estimate between 2 sample groups.
In a t-test the degree of freedom is calculated as the total sample size minus 1 i.e. $df = \sum ns - 1$, where “ n_s ” is the number of observations in the sample. Suppose, we have 2 samples A and B. The df would be calculated as $df = (nA - 1) + (nB - 1)$
- **Significance Level:** The significance level is the predetermined threshold that is used to decide whether to reject the null hypothesis. Commonly used significance levels are 0.05, 0.01, or 0.10.
- **T-statistic:** The t-statistic is a measure of the difference between the means of two groups. It is calculated as the difference between the sample means divided by the standard error of the difference. It is also known as the t-value or t-score.
 - If the t-value is large => the two groups belong to different groups.
 - If the t-value is small => the two groups belong to the same group.

T-Statistics

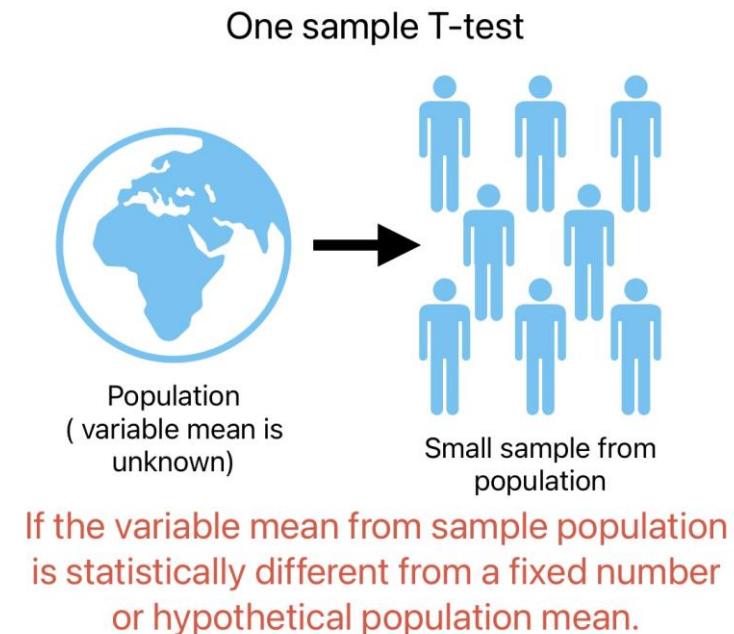


One Sample T-Test

The value against which we are comparing is a single value, i.e. we compare the mean of a sample with a single value to check how much the mean deviates from that single value.

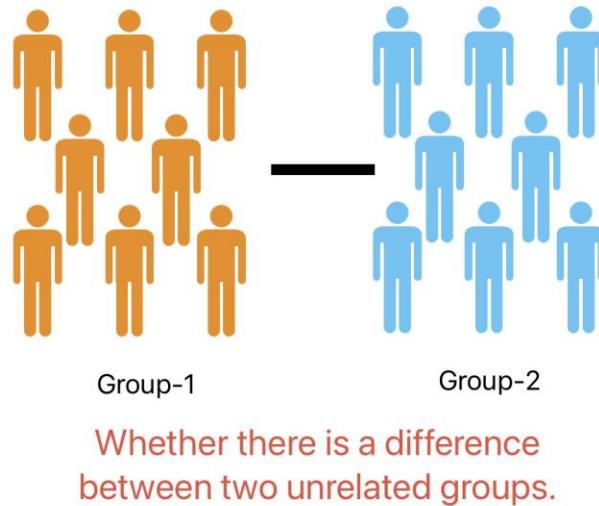
Two Sample T-Test

We compare the means and variances of two samples, we assess how much they differ.



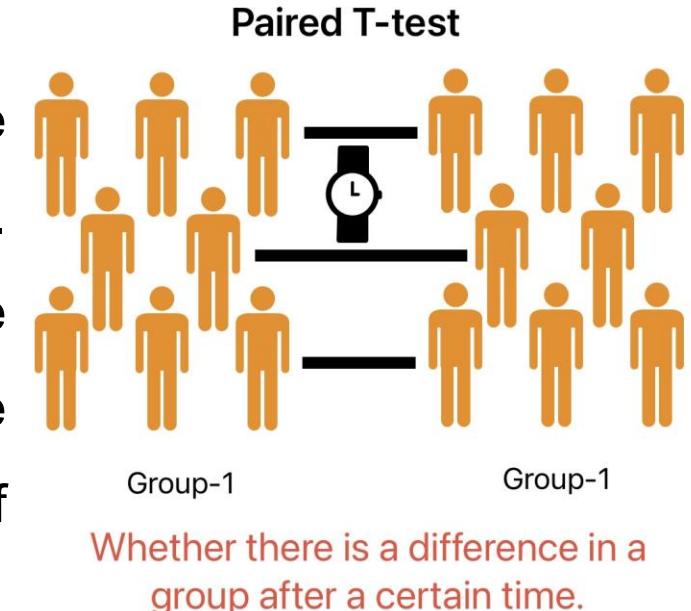
T-Statistics

Independent two sample T-test
(Unpaired t-test)



In an **independent two-sample t-test** (unpaired t-test), the samples in the two groups being compared are unrelated. The samples are drawn from two different populations or groups of subjects, and the difference between the means of the two groups is calculated using the means and variances of the two separate samples.

In a **dependent two-sample t-test** (also known as a **paired t-test**), the samples in the two groups being compared are related in some way. For example, the samples may be pairs of measurements taken on the same subjects. In this case, the difference between the means of the two groups is calculated by taking the differences between the pairs of measurements and treating these differences as a single sample.



T-Statistics

One-Sample T-Test

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

\bar{x} = observed mean of the sample
 μ = assumed mean
 s = standard deviation
 n = sample size

Two-Sample T-Test

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

\bar{x}_1 = observed mean of 1st sample
 \bar{x}_2 = observed mean of 2nd sample
 s_1 = standard deviation of 1st sample
 s_2 = standard deviation of 2nd sample
 n_1 = sample size of 1st sample
 n_2 = sample size of 2nd sample

Paired sample T-test

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

where: $s_d = \text{sqrt}[\sum (d_i - \bar{d})^2 / (n - 1)]$

- d is the mean of the difference scores
- s_d is the standard deviation of the difference scores
- n is the number of pairs of observations

d_i : The difference between the paired measurements for the i -th participant (After - Before for the i -th participant)
 \bar{d} : The mean difference (average of all the d_i 's)

T-Statistics

Q1. A manufacturer claims that the average weight of their product is 50 grams. You want to test this claim. You randomly sample 25 products and find the following weights (in grams):

48, 52, 49, 51, 50, 47, 53, 50, 49, 52, 48, 51, 50, 49, 50, 51, 48, 52, 49, 50, 51, 47, 53, 50, 49

1. State the hypotheses:

- Null hypothesis (H_0): The population mean is equal to 50 grams ($\mu = 50$).
- Alternative hypothesis (H_a): The population mean is not equal to 50 grams ($\mu \neq 50$).

2. Calculate the sample mean (\bar{x}) and sample standard deviation (s):

- $\bar{x} = (\text{sum of all weights}) / (\text{number of samples}) = 1245 / 25 = 49.8$ grams
- $s = (\text{square root of } [\text{sum of } (\text{each weight} - \bar{x})^2] / (\text{number of samples} - 1)) \approx 1.83$ grams

3. Calculate the t-statistic:

- $t = (\bar{x} - \mu) / (s / \sqrt{n}) = (49.8 - 50) / (1.83 / \sqrt{25}) \approx -0.55$

4. Determine the degrees of freedom (df):

- $df = n - 1 = 25 - 1 = 24$

5. Find the critical value:

- You need to choose a significance level (alpha). Let's say $\alpha = 0.05$.
- Consult a t-distribution table or use a calculator to find the critical value for a two-tailed test with $df = 24$ and $\alpha = 0.05$. The critical value is approximately ± 2.064 .

Since our calculated t-statistic (-0.55) falls within the range of -2.064 to +2.064, we fail to reject the null hypothesis.

Conclusion: There is not enough evidence to conclude that the average weight of the products is different from 50 grams.

t-test table

T-Statistics

Q2. A researcher wants to know whether there is a significant difference in the average test scores of students who are taught using two different methods. The researcher randomly assigns 20 students to one of two groups. Group A is taught using method A, and group B is taught using method B. After the students have completed the course, they are given a test. The test scores are shown below:

Group A: 85, 90, 92, 88, 89, 91, 93, 87, 86, 94

Group B: 78, 82, 80, 85, 83, 81, 79, 84, 77, 86

Solution:

- t-test is used to determine whether there is a significant difference in the average test scores of the two groups.
- The null hypothesis is that there is no significant difference in the average test scores of the two groups.
- The alternative hypothesis is that there is a significant difference in the average test scores of the two groups.
- Calculates the mean and standard deviation of the test scores for each group.

Group A: mean = 89.5, standard deviation = 2.87

Group B: mean = 81.5, standard deviation = 2.87

T-Statistics

- The null hypothesis is that there is no significant difference in the average test scores of the two groups.
- The alternative hypothesis is that there is a significant difference in the average test scores of the two groups.
- Calculates the mean and standard deviation of the test scores for each group.

Group A: mean = 89.5, standard deviation = 2.87

Group B: mean = 81.5, standard deviation = 2.87

- Calculate t-score using two sample t-test formula:

$$t = (89.5 - 81.5) / \sqrt{ (2.87^2 / 10) + (2.87^2 / 10) } = 5.57$$

- Degree of freedom df = $= (nA - 1) + (nB - 1) = 18$

Let, the significance level is 0.05, the critical value is obtained from the t-table as 2.101.

Since the t-statistic (5.57) is greater than the critical value (2.101), the researcher rejects the null hypothesis. This means that there is a significant difference in the average test scores of the two groups.

t-test table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

T-Statistics

Q3. A researcher wants to study the effectiveness of a new memory-enhancing drug. They recruit 20 participants and administer a memory test before and after taking the drug for a month. Is there a statistically significant difference in memory test scores before and after taking the drug?

Solution: this is a paired t-test question, as we have two measurements (before and after) for each participant.

1. Calculate the Differences: Subtract the "Before Drug" score from the "After Drug" score for each participant.

2. Calculate the Mean Difference (\bar{d}): Sum the differences and divide by the number of participants ($n = 20$).

$$\bar{d} = (7+7+5+6+5+2+7+5+5+7+5+6+6+4+4+7+5+5+6) / 20 = 5.35$$

3. Calculate the Standard Deviation of the Differences (sd):

$$sd = \sqrt{\sum (di - \bar{d})^2 / (n - 1)} = 1.35$$

4. Calculate the t-statistic: $t = (\bar{d}) / (sd / \sqrt{n}) = 5.35 / (1.35 / \sqrt{20}) \approx 17.7$

5. Determine the Degrees of Freedom: $df = n - 1 = 20 - 1 = 19$

6. Find the Critical Value: for alpha 0.05 and df = 19, critical value is ± 2.093 . Since the absolute value of our t-statistic is greater than the critical value, we reject the null hypothesis.

Participant	Before Drug	After Drug	Difference (After - Before)
1	75	82	7
2	68	75	7
3	80	85	5
4	72	78	6
5	85	90	5
6	70	72	2
7	78	85	7
8	65	70	5
9	90	95	5
10	75	80	5
11	72	79	7
12	68	73	5
13	82	88	6
14	77	83	6
15	88	92	4
16	71	75	4
17	79	86	7
18	66	71	5
19	84	89	5
20	73	79	6

Statistics

Chi-Square test

In recent years, the use of specialized statistical methods for categorical data has increased dramatically, particularly for applications in the biomedical and social sciences. Categorical scales occur frequently in the health sciences, for measuring responses. E.g.

- patient survives an operation (yes, no),
- severity of an injury (none, mild, moderate, severe), and
- stage of a disease (initial, advanced).

Studies often collect data on categorical variables that can be summarized as a series of counts and commonly arranged in a tabular format known as a **contingency table**.

Chi-Square test χ^2

The most obvious difference between the chi-square tests and the other hypothesis tests we have considered (T test) is the nature of the data (categorical data).

- For chi-square, the data are **frequencies** rather than numerical scores.
- Used for testing significance of patterns in qualitative data.
- Test statistic is based on counts (frequencies) that represent the number of items that fall in each category
- Test statistics measures the agreement between actual counts(observed) and expected counts assuming the null hypothesis

Chi-Square test χ^2

Chi-square Test – Distribution table and formulas

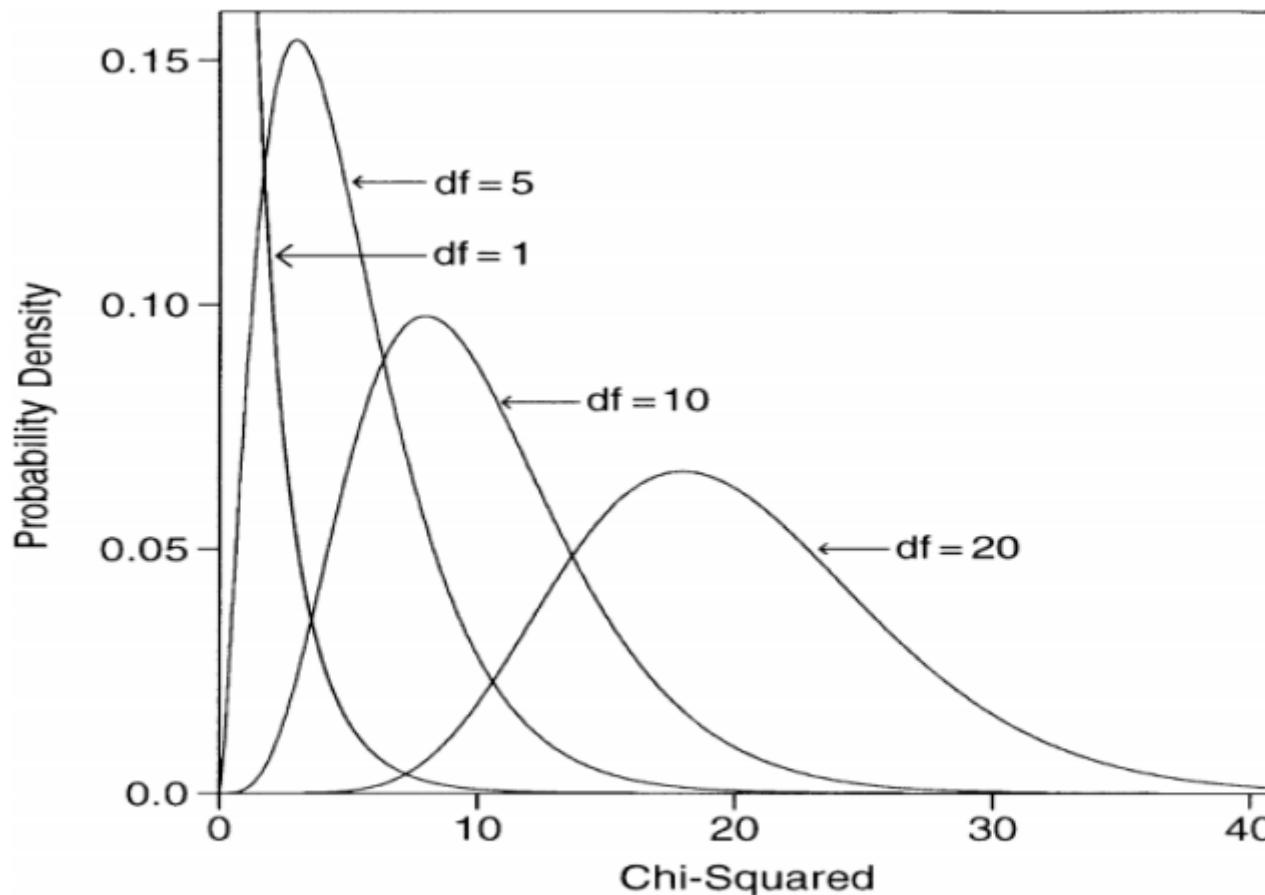
Degrees of Freedom (df) Significance Level (α)	0.01	0.05	0.10	0.25	0.50
1	6.635	3.841	2.706	1.323	0.454
2	9.210	5.991	4.605	2.773	1.386
3	11.345	7.815	6.251	3.930	2.366
4	13.277	9.488	7.779	5.178	3.357
5	15.086	11.070	9.236	6.571	4.351
6	16.812	12.592	10.645	7.962	5.348
7	18.475	14.067	12.017	9.364	6.346
8	20.090	15.507	13.362	10.773	7.344
9	21.666	16.919	14.684	12.189	8.343
10	23.209	18.307	15.987	13.603	9.342

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$df = (r-1) \times (c-1)$$

$$E_i = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

Chi-Square test χ^2



- The degrees of freedom for tests of hypothesis that involve an rxc contingency table is **equal to $(r-1) \times (c-1)$** ;

Chi-Square test χ^2

Application of chi square test

1. **Goodness-of-fit:** uses frequency data from a sample to test hypotheses about the shape or proportions of a population.
2. **Test for independence:**
 1. (2×2 chi-square test): Testing hypotheses about the relationship between two variables in a population,
 2. ($a \times b$ chi-square test) or ($r \times c$ chi-square test)

Chi-Square test χ^2

Q1. Given Eye colour in a sample of 40 people: Blue 12, brown 21, green 3, others 4

Given Eye colour in population: Brown 80%, Blue 10%, Green 2%, Others 8%

Is there any difference between proportion of sample to that of population (use alpha= 0.05)

Solution: Assume Sample is randomly selected from the population.

Null hypothesis: there is no significant difference in proportion of eye colour of sample to that of the population.

Alternative hypothesis: there is significant difference in proportion of eye colour of sample to that of the population.

$$\begin{aligned}\chi^2 &= \frac{(12-4)^2}{4} + \frac{(21-32)^2}{32} + \frac{(3-0.8)^2}{0.8} + \frac{(4-3)^2}{3} \\ &= (64/4) + (121/32) + (4.8/0.8) + (1/3) \\ &= 16 + 3.78 + 6 + 0.3 \\ &= 26.08\end{aligned}$$

Color	Sample frequency	Expected frequency
Blue	12	$40*10/100= 4$
Brown	21	$40*80/100=32$
Green	3	$40*2/100=0.8$
Others	4	$40*8/100=3$

Chi-Square test χ^2

Q1. Given Eye colour in a sample of 40 people: Blue 12, brown 21, green 3, others 4

Given Eye colour in population: Brown 80%, Blue 10%, Green 2%, Others 8%

Is there any difference between proportion of sample to that of population (use alpha= 0.05)

Solution: Assume Sample is randomly selected from the population.

Null hypothesis: there is no significant difference in proportion of eye colour of sample to that of the population.

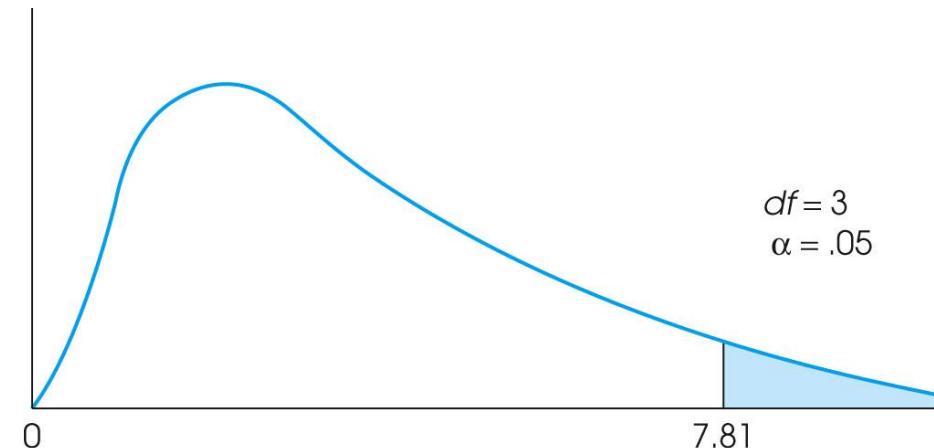
Alternative hypothesis: there is significant difference in proportion of eye colour of sample to that of the population.

$$\alpha = 0.05$$

$$d.f.(\text{degree of freedom}) = K - 1 = 4 - 1 = 3$$

(K=Number of subgroups)

critical value for $\alpha = 0.05$ and $df = 3 \Rightarrow 7.81$

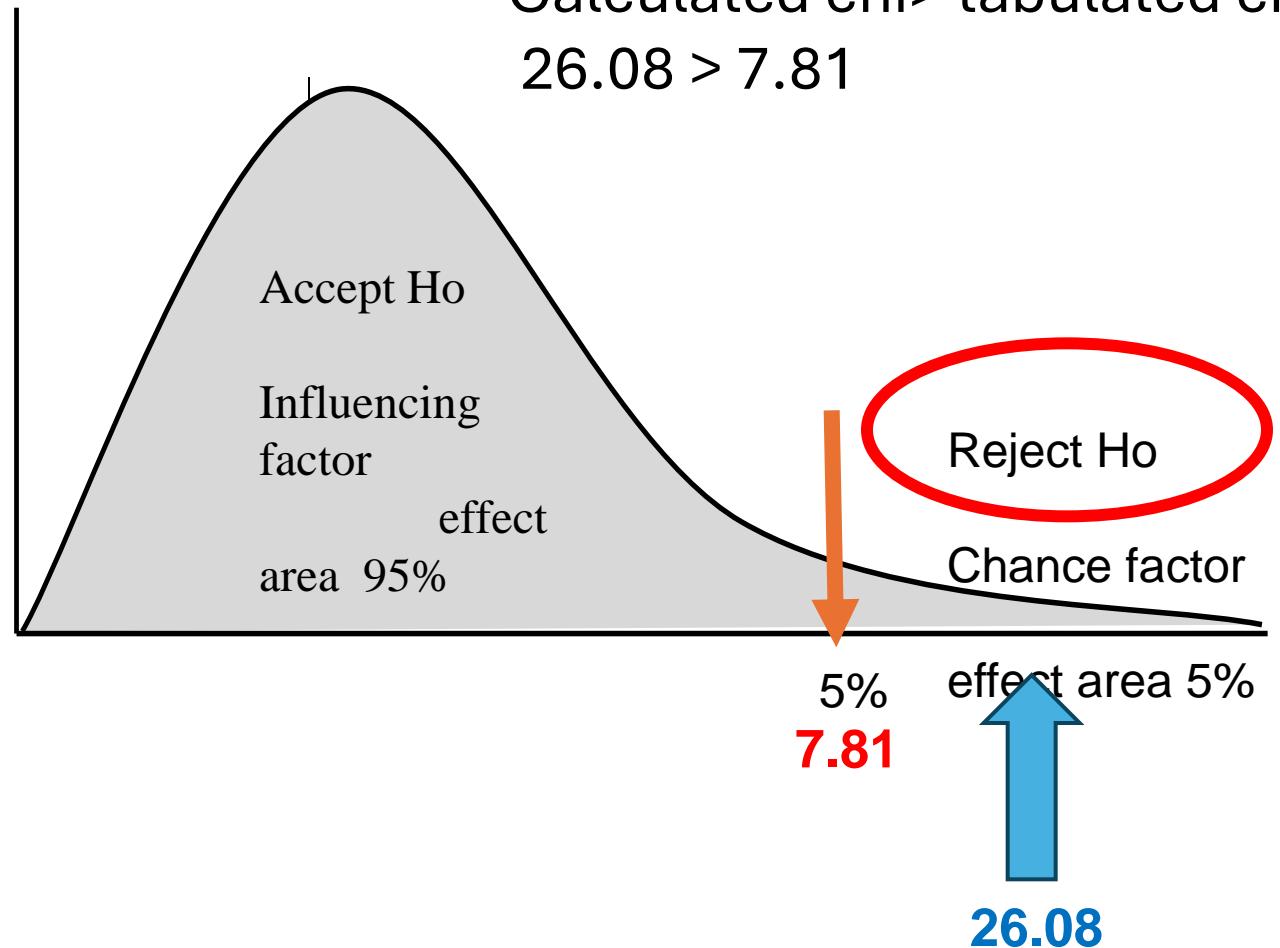


Chi-Square test

χ^2

Conclusion: We reject H_0 & accept H_A
There is significant difference in
proportion of eye colour of sample to
that of the population.

Calculated chi > tabulated chi
 $26.08 > 7.81$



Chi-Square test χ^2

Q2. A total 1500 workers on 2 operators (A&B) were classified as deaf & non-deaf according to the following table. Is there association (dependence) between deafness & type of operator. Let $\alpha = 0.05$

HO: there is no significant **association** between type of operator & deafness.

HA: there is significant **association** between type of operator & deafness.

$\alpha = 0.05$

d.f.(degree of freedom)=(2-1)(2-1) = 1

critical value for $\alpha = 0.05$ and df=1 => 3.841

Operator	deaf	Not deaf.	total
A	100	900	1000
B	60	440	500
total	160	1340	1500

Total number of items=1500

Total number of defective items=160

Chi-Square test χ^2

Q2. A total 1500 workers on 2 operators (A&B) were classified as deaf & non-deaf according to the following table. Is there association (dependence) between deafness & type of operator. Let $\alpha = 0.05$

$$\text{Expected deaf from Operator A} = 1000 * 160/1500 = 106.7$$

$$(\text{expected not deaf} = 1000 - 106.7 = 893.3)$$

$$\text{Expected deaf from Operator B} = 500 * 160/1500 = 53.3$$

$$\begin{aligned}\chi^2 &= \frac{(100-106.7)^2}{106.7} + \frac{(900-893.3)^2}{893.3} + \frac{(60-53.3)^2}{53.3} + \frac{(440-446.7)^2}{446.7} \\ &= 0.42 + 0.05 + 0.84 + 0.10 = 1.41\end{aligned}$$

Operator	deaf	Not deaf.	total
A	100	900	1000
B	60	440	500
total	160	1340	1500

Total number of items = 1500

Total number of defective items = 160

Chi-Square test χ^2

Q2. A total 1500 workers on 2 operators (A&B) were classified as deaf & non-deaf according to the following table. Is there association (dependence) between deafness & type of operator. Let $\alpha = 0.05$

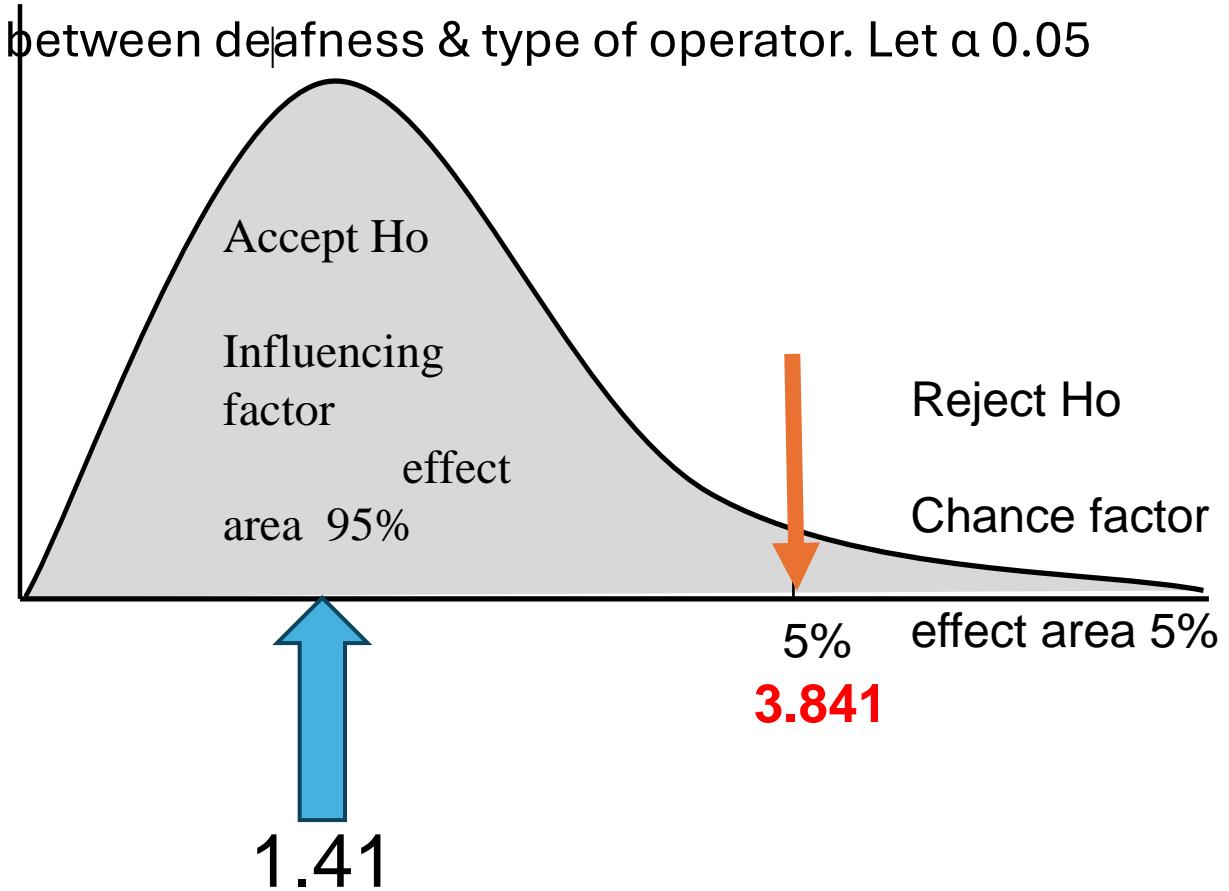
Calculated chi < tabulated chi

$$1.41 < 3.841$$

Conclusion: We accept H_0

H_0 may be true

There is no significant association between type of operator & deafness.



Chi-Square test

**Test for Independence using
(a x b chi-square test) or
(r x c chi-square test)**

Calculation of expected frequencies: For $r \times c$ contingency table, the expected frequencies are as follow:

$$e_i = \frac{\text{Row total}(rt_i) \times \text{Column total}(ct_i)}{\text{Grand total}(n)}$$

Where e_i = expected frequency of cells and is e_1, e_2, \dots, e_k where k is the number of cells in the body of the table.

Consider the following 3 by 2 contingency table

	<i>Classification criteria 2</i>	<i>Classification criteria 1</i>		
		<i>Class 1</i>	<i>Class 2</i>	<i>Total</i>
Category 1		a	b	$a + b$
Category 2		c	d	$c + d$
Category 3		e	f	$e + f$
<i>Total</i>		$a + c + e$	$b + d + f$	n

The expected value for the first cell (a), $e_1 = \frac{(a+b)(a+c+e)}{n}$

The expected value for the first cell (b), $e_2 = \frac{(a+b)(b+d+f)}{n}$

.....
The expected value for the first cell (f), $e_6 = \frac{(e+f)(b+d+f)}{n}$

Chi-Square test χ^2

Q3. Perform a Chi-Square test to analyze the relationship between alcohol consumption (number of beers per day) and liver disease. The contingency table is given below.

1. State the Hypotheses:

Null Hypothesis (H_0): There is no association between the number of beers consumed per day and the presence of liver disease. The two variables are independent.

Alternative Hypothesis (H_1): There is an association between the number of beers consumed per day and the presence of liver disease. The two variables are not independent.

2. The expected frequency for each cell is calculated as:

$$(\text{Row Total} * \text{Column Total}) / \text{Grand Total}$$

Expected values are shown in brackets with each cell.

3. Calculate the Chi-Square Statistic (χ^2):

$$\chi^2 = \sum [(\text{Observed} - \text{Expected})^2 / \text{Expected}]$$

$$= 35.71 + 83.33 + 0.43 + 1.00 + 2.21 + 7.74 = 153.4$$

4. Find critical value for $df = (3-1)(2-1) = 2$ and alpha = 0.05

Critical value = 5.991

5. Compare: Our calculated χ^2 (130.42) is much greater than the critical value (5.991). Therefore, we reject the null hypothesis.

<i>Alcohol Drinking (No. of bottle beers/day)</i>	<i>Liver Disease</i>		<i>Total</i>
	<i>Yes</i>	<i>No</i>	
≤ 2	20	80	100
3-5	90	30	120
≥ 6	240	40	280
Total	350	150	500

<i>Beers/Day</i>	<i>Liver Disease (Yes)</i>	<i>Liver Disease (No)</i>	<i>Total</i>
≤ 2	20 (70)	80 (30)	100
3-5	90 (84)	30 (36)	120
≥ 6	240 (196)	40 (84)	280
Total	350	150	500

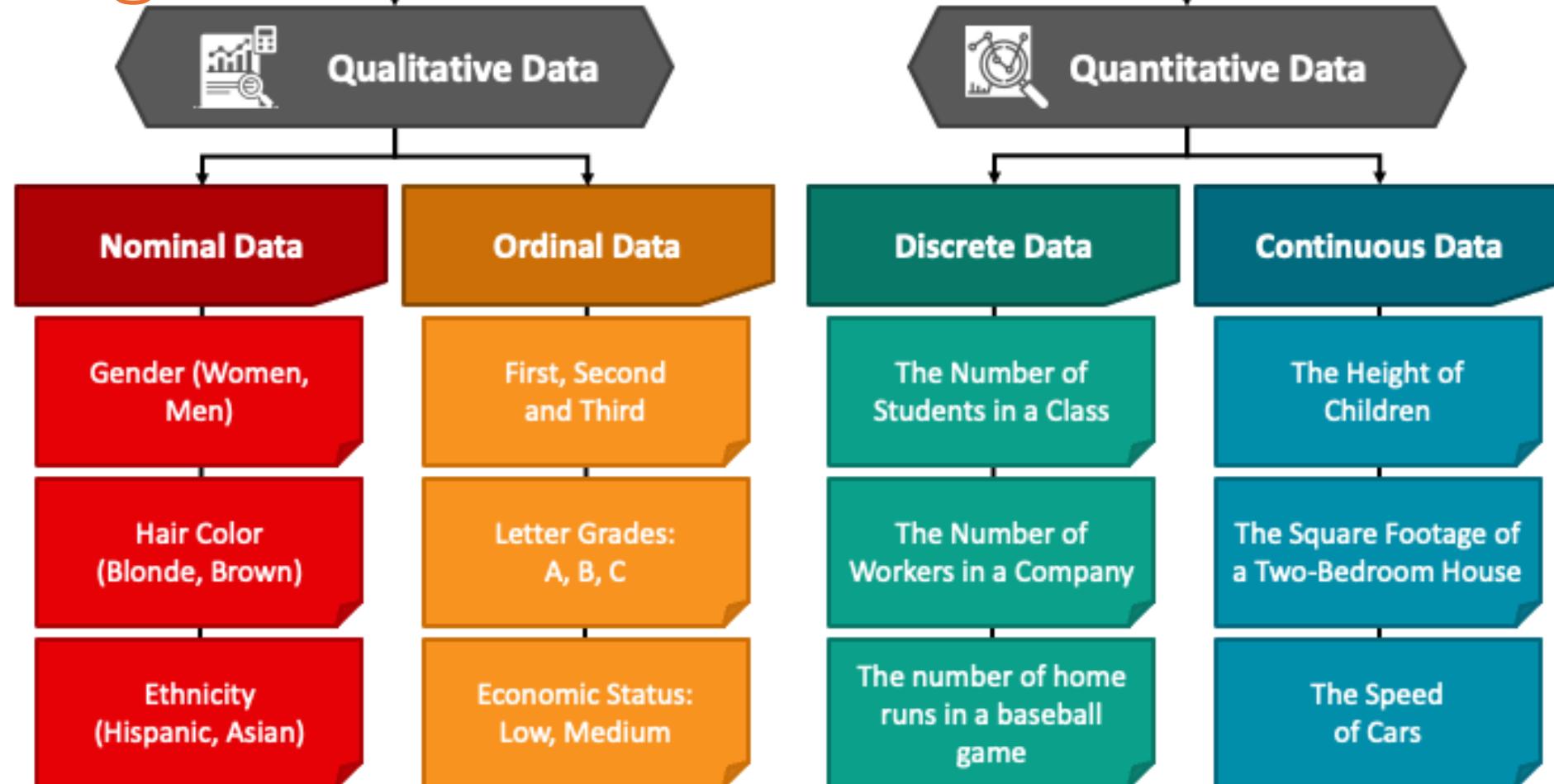
Conclusion: There is a statistically significant association between the number of beers consumed per day and the presence of liver disease. The variables are not independent.

Data and Data Sources

Categorical

TYPES OF DATA

Numerical



Categorical Data

- The objects being studied are grouped into categories based on some **qualitative** trait.
- The resulting data are merely labels or categories.
- E.g. Hair color (blonde, brown, red, black, etc.), Opinion of students about riots (ticked off, neutral, happy), Smoking status (smoker, non-smoker), etc.
- **Nominal:** A type of categorical data in which objects fall into **unordered** categories.
 - E.g. Hair color (blonde, brown, red, black, etc.), Smoking status (smoker, non-smoker), etc.
- **Ordinal:** A type of categorical data in which **order** is important.
 - E.g. Class (fresh, sophomore, junior, senior, super senior), Opinion of students about riots (ticked off, neutral, happy), etc.
- **Binary:** A type of categorical data in which there are only two categories. Can be nominal or ordinal.
 - E.g. Attendance (present, absent), Smoking status (smoker, non-smoker), etc.

Numerical

- The objects being studied are “measured” based on some **quantitative** trait.
- The resulting data are set of numbers.
- E.g., Cholesterol level, Height, Age, SAT score, Number of students late for class, Time to complete a homework assignment, etc.
- **Discrete:** Only certain values are possible (there are gaps between the possible values).
 - E.g., SAT scores, Number of students late for class, etc.
- **Continuous:** Theoretically, any value within an interval is possible with a fine enough measuring device.
 - E.g., Cholesterol level, Height, Time to complete a homework assignment, etc.

Types of Data

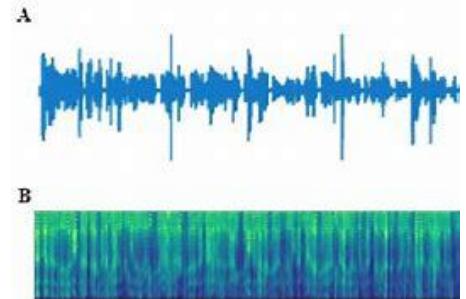
- The type(s) of data collected in a study determine the type of statistical analysis used. For instance:
- Categorical data are commonly summarized using “**percentages**” (or “**proportions**”).
 - 11% of students have a tattoo
 - 2%, 33%, 39%, and 26% of the students in class are, respectively, freshmen, sophomores, juniors, and seniors
- Numerical data are typically summarized using “**averages**” (or “**means**”).
 - Average number of siblings Fall 1998 Stat 250 students have is 1.9.
 - Average weight of male Fall 1998 Stat 250 students is 173 pounds.
 - Average weight of female Fall 1998 Stat 250 students is 138 pounds.

Types of Data

Based on Structure

- **Structured Data** – Organized, stored in relational databases (e.g., SQL tables).
 - Example: Customer records, transaction logs.
- **Unstructured Data** – No predefined format, difficult to analyze (e.g., text, images, videos).
 - Example: Social media posts, raw audio files.
- **Semi-Structured Data** – Some structure but not strictly formatted (e.g., JSON, XML).
 - Example: Emails, NoSQL databases.

A	B	C	D	E
Cookie Sales by Region				
1	SalesRep	Region	# Orders	Total Sales
2	Bill	West	217	\$41,107
3	Frank	West	268	\$72,707
4	Harry	North	224	\$41,676
5	Janet	North	286	\$87,858
6	Joe	South	226	\$45,606
7	Martha	East	228	\$49,017
8	Mary	West	234	\$57,967
9	Ralph	East	267	\$70,702
10	Sam	East	279	\$77,738
11	Tom	South	261	\$69,496
12				
13				
14				
15				



```
{"widget": {  
    "debug": "on",  
    "window": [  
        {"title": "Sample Konfabulator Widget",  
         "name": "main_window",  
         "width": 500,  
         "height": 500  
     },  
     {"image": {  
        "src": "Images/Sun.png",  
        "name": "sun1",  
        "xOffset": 250,  
        "yOffset": 250,  
        "alignment": "center"  
    }  
},  
    "text": {  
        "data": "Click Here",  
        "size": 36,  
        "style": "bold",  
        "name": "text1",  
        "xOffset": 250,  
        "yOffset": 100,  
        "alignment": "center",  
        "onMouseUp": "sun1.opacity = (sun1.opacity / 100) + 90;"  
    }  
}}}
```

Types of Data

Text Data

- **Definition:** Data in the form of natural language text.
- **Examples:** Social media posts, customer reviews, research papers.
- **Use Cases:** Sentiment analysis, language translation, text classification.

Image Data

- **Definition:** Data in the form of images.
- **Examples:** Photographs, medical scans, satellite images.
- **Use Cases:** Image classification, object detection, image segmentation.

Audio Data

- **Definition:** Data in the form of sound recordings.
- **Examples:** Speech recordings, music, environmental sounds.
- **Use Cases:** Speech recognition, audio classification, music generation.

Video Data

- **Definition:** Data in the form of moving images.
- **Examples:** Surveillance footage, video clips, movies.
- **Use Cases:** Action recognition, video summarization, video segmentation.

Types of Datasets

Datasets can be categorized based on their characteristics, structure, and the type of problem they are used to solve. Different types of datasets serve various purposes and are used in different applications. Understanding the nature of the data and choosing the right type of dataset for a specific problem is crucial for developing effective and accurate prediction models.

Here are the primary types of datasets:

1. Structured vs. Unstructured Dataset

Structured Data

- **Definition:** Data that is organized in a predefined manner, often in tabular format with rows and columns.
- **Examples:** Spreadsheets, SQL databases.
- **Use Cases:** Financial records, customer databases, sensor data.

Unstructured Data

- **Definition:** Data that does not have a predefined format or organization.
- **Examples:** Text documents, images, audio files, videos.
- **Use Cases:** Natural language processing (NLP), image recognition, speech-to-text conversion.

Types of Datasets

2. Labeled vs. Unlabeled Datasets

Labeled Data

- **Definition:** Data that has been tagged with one or more labels, providing explicit information about the target variable.
- **Examples:** Annotated images (with objects labeled), spam vs. non-spam emails.
- **Use Cases:** Supervised learning tasks such as classification and regression.

Unlabeled Data

- **Definition:** Data without any labels or target variables.
- **Examples:** Raw text, unlabeled images, customer behavior data.
- **Use Cases:** Unsupervised learning tasks such as clustering, anomaly detection.

3. Time Series Datasets

- **Definition:** Data points collected or recorded at specific time intervals.
- **Examples:** Stock prices, weather data, sensor readings.
- **Use Cases:** Forecasting, anomaly detection, trend analysis.

Types of Datasets

4. Training, Validation, and Test Sets

Training Set

- **Definition:** The portion of the dataset used to train the machine learning model.
- **Purpose:** To allow the model to learn patterns and relationships in the data.

Validation Set

- **Definition:** A subset of the dataset used to tune model parameters and make decisions about model architecture.
- **Purpose:** To provide an unbiased evaluation of a model fit on the training dataset while tuning hyperparameters.

Test Set

- **Definition:** The portion of the dataset used to evaluate the final model performance.
- **Purpose:** To provide an unbiased assessment of the model's performance on unseen data.

Popular sources for datasets

- <https://www.kaggle.com/>
- [Hugging Face – The AI community building the future.](#)
- [Home - UCI Machine Learning Repository](#)
- [Dataset Storage and Dataset Search Platform | IEEE DataPort](#)

Data Quality and Issues

Data quality is crucial for accurate analysis and decision-making. . Poor data quality leads to incorrect insights and business risks.

Key Dimensions of Data Quality

Accuracy – Data should be correct and free from errors.

Completeness – No missing values or gaps in data.

Consistency – Uniform format and values across different sources.

Timeliness – Data should be up to date and relevant.

Validity – Data should conform to predefined formats and rules.

Uniqueness – No duplicate or redundant records.

Data Quality and Issues

Common Data Issues

- **Missing Data** – Incomplete records causing bias in analysis.
- **Duplicate Data** – Multiple records for the same entity.
- **Inconsistencies** – Data mismatch across different systems.
- **Incorrect Data** – Human or system errors in data entry.
- **Data Drift** – Changes in data patterns over time affecting model accuracy.
- **Bias in Data** – Unrepresentative data leading to skewed results.
- **Scalability:** Handling large volumes of data requires scalable storage and processing solutions.

Data Quality and Issues

Improving Data Quality

- **Data Cleaning** – Removing errors, duplicates, and inconsistencies.
- **Data Standardization** – Ensuring a uniform format across datasets.
- **Validation Techniques** – Implementing rule-based validation and automated error detection.
- **Data Governance** – Setting policies for data integrity, security, and compliance.
- **Continuous Monitoring** – Regular data audits and updates to maintain quality.

Association Analysis and Prediction Analysis

Feature	Association Analysis	Prediction Analysis
Goal	To discover relationships and patterns between variables.	To build models that forecast future outcomes or unknown values.
Focus	Identifying "what goes with what."	Predicting "what will happen."
Purpose	Understanding the relationships and dependencies within data.	Maximizing the accuracy of predictions on new data.
Output	Rules, correlations, or patterns describing relationships.	A predictive model that produces forecasts.
Evaluation	Statistical significance of relationships (e.g., support, confidence, lift).	Accuracy metrics (e.g., precision, recall, accuracy, RMSE).
Interpretability	Often high; relationships are typically easier to understand.	Varies; complex models may be "black boxes" with low interpretability.
Example	Market basket analysis (finding which items are frequently bought together).	Predicting customer churn or forecasting sales.
Key Question	"What variables are related?"	"What outcome is most likely?"
Typical Methods	Association rule mining, correlation analysis.	Regression analysis, classification algorithms, time series analysis.

Data and Data Sources

Market Basket analysis



- It identifies associations between products in transactions.
- Uses Association Rule Mining to generate rules like "If a customer buys X, they are likely to buy Y."
- Commonly applied in retail, e-commerce, and recommendation systems.

Association Analysis

Association Analysis is a data mining technique used to **discover relationships or patterns** between items in large datasets. It is widely used in **market basket analysis, recommendation systems, fraud detection, and web usage mining**.

Objective:

To find frequent item-sets and association rules that describe how items are related within a dataset.

Association Rules

Association rules are statements in the form of:

If $X \Rightarrow Y$, Which means, **if item X appears, item Y is also likely to appear**.

Association Analysis

Example:

- **{Bread, Butter} → {Milk}** (People who buy bread and butter often buy milk)
- **{Laptop} → {Mouse}** (People who buy a laptop are likely to buy a mouse)

Key metrics used to evaluate association rules:

1. Support

Measures how frequently an itemset appears in the dataset.

$$Support(X) = \frac{\text{Frequency of } X \text{ in dataset}}{\text{Total transactions}}$$

Association Analysis

Example: If Milk appears in 30 out of 100 transactions, then:

$$Support(Milk) = \frac{30}{100} = 30\%$$

2. Confidence

Measures how often Y appears when X is present.

$$Confidence(X \Rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)}$$

Example: If Bread appears in 50 transactions, and in 40 of them, Milk is also bought:

$$Confidence(Bread \Rightarrow Milk) = \frac{40}{50} = 80\%$$

Association Analysis

3. Lift

Measures how much **stronger** the association is compared to a random occurrence.

$$Lift(X \Rightarrow Y) = \frac{\text{Confidence}(X \rightarrow Y)}{\text{Support}(Y)}$$

If $Lift > 1$: X and Y are positively correlated (buying one increases the likelihood of buying the other).

If $Lift < 1$: X and Y are negatively correlated (buying one reduces the likelihood of buying the other).

Problems

A retailer wants to analyze buying patterns based on 500 transactions in a week:

- $\{\text{Laptop}\}$ appears in 100 transactions.
- $\{\text{Laptop}, \text{Mouse}\}$ together appear in 60 transactions.
- $\{\text{Mouse}\}$ appears in 150 transactions.

Questions:

1. What is the confidence of the rule $\{\text{Laptop}\} \rightarrow \{\text{Mouse}\}$?
2. What is the confidence of the rule $\{\text{Mouse}\} \rightarrow \{\text{Laptop}\}$?

Problems: Supermarket Transactions

Transaction Dataset

Transaction ID	Items Purchased
T1	Milk, Bread, Butter
T2	Bread, Butter
T3	Milk, Bread
T4	Milk, Bread, Butter, Eggs
T5	Bread, Butter, Eggs

Step 1: Compute Support

- Support(Milk)
- Support(Bread)
- Support(Butter)
- Support({Milk, Bread})
- Support({Bread, Butter})

Step 2: Compute Confidence

- Confidence(Milk → Bread)
- Confidence(Bread → Butter)

Step 3: Compute Lift

- Lift(Milk → Bread)
- Lift(Bread → Butter)

Problems

Transaction Data

Transaction ID	Items Purchased
T1	Apple, Banana, Milk
T2	Apple, Banana
T3	Apple, Banana, Milk
T4	Banana, Milk, Bread
T5	Apple, Bread
T6	Banana, Bread
T7	Apple, Banana, Bread

Lift(Apple → Banana)

0.87

Lift(Banana → Bread)

0.60

Applications of Market Basket analysis

Retail:

- Optimize product placement (e.g., placing Milk near Bread).
- Identify frequently bought-together items for promotions.

E-commerce & Recommendations:

- Suggest items frequently bought together (Amazon's "Customers who bought this also bought...").
- Improve personalized recommendations.

Healthcare: Analyze patient symptoms and medications that are frequently prescribed together.

Finance: Detect fraud by identifying unusual spending patterns.

Practice

Q. Using the following transactional dataset of customer purchases. Find:

- i. Frequent Itemset/s
- ii. Association rules
- iii. Support, confidence and lift of the rules

Transaction ID	Items Purchased
1	Bread, Milk, Eggs
2	Bread, Butter
3	Milk, Butter
4	Bread, Milk, Butter, Cheese
5	Eggs, Milk
6	Bread, Eggs
7	Milk
8	Bread, Butter, Milk

Practice

Frequent Itemsets (Let's use a minimum support of 2):

- Individual Items:

- Bread: $5/8 = 0.625$ (Support = 0.625)
- Milk: $6/8 = 0.75$ (Support = 0.75)
- Eggs: $3/8 = 0.375$ (Support = 0.375)
- Butter: $4/8 = 0.5$ (Support = 0.5)
- Cheese: $1/8 = 0.125$ (Support = 0.125)

- Pairs:

- {Bread, Milk}: $3/8 = 0.375$ (Support = 0.375)
- {Bread, Butter}: $3/8 = 0.375$ (Support = 0.375)
- {Milk, Butter}: $3/8 = 0.375$ (Support = 0.375)
- {Milk, Eggs}: $2/8 = 0.25$ (Support = 0.25)
- {Bread, Eggs}: $2/8 = 0.25$ (Support = 0.25)

- Triplets:

- {Bread, Milk, Butter}: $2/8 = 0.25$ (Support = 0.25)

Transaction ID	Items Purchased
1	Bread, Milk, Eggs
2	Bread, Butter
3	Milk, Butter
4	Bread, Milk, Butter, Cheese
5	Eggs, Milk
6	Bread, Eggs
7	Milk
8	Bread, Butter, Milk

Practice

Support, Confidence, and Lift:

- **Support:** The proportion of transactions that contain the itemset.
- **Confidence:** The probability that a transaction containing A also contains B ($A \rightarrow B$).
- **Lift:** The ratio of the observed support to the support if A and B were independent. A lift greater than 1 suggests a positive association.

Association Rules (Using the frequent itemsets):

• **{Bread, Milk} \rightarrow {Butter}:**

$$\text{Support} = 2/8 = 0.25$$

$$\text{Confidence} = 2/3 = 0.666$$

$$\text{Lift} = (2/8) / ((3/8) * (4/8)) = 1.33$$

• **{Bread, Butter} \rightarrow {Milk}:**

$$\text{Support} = 2/8 = 0.25$$

$$\text{Confidence} = 2/3 = 0.666$$

$$\text{Lift} = (2/8) / ((3/8) * (6/8)) = 0.888$$

• **{Milk, Butter} \rightarrow {Bread}:**

$$\text{Support} = 2/8 = 0.25$$

$$\text{Confidence} = 2/3 = 0.666$$

$$\text{Lift} = (2/8) / ((3/8) * (5/8)) = 1.066$$

Transaction ID	Items Purchased
1	Bread, Milk, Eggs
2	Bread, Butter
3	Milk, Butter
4	Bread, Milk, Butter, Cheese
5	Eggs, Milk
6	Bread, Eggs
7	Milk
8	Bread, Butter, Milk

Practice

Transaction ID	Items Purchased
1	Bread, Milk, Eggs
2	Bread, Butter
3	Milk, Butter
4	Bread, Milk, Butter, Cheese
5	Eggs, Milk
6	Bread, Eggs
7	Milk
8	Bread, Butter, Milk

Rule	Support	Confidence	Lift
{Bread} -> {Milk}	0.375	0.6	0.8
{Bread} -> {Butter}	0.375	0.6	1.2
{Milk} -> {Bread}	0.375	0.5	0.8
{Milk} -> {Butter}	0.375	0.5	1
{Butter} -> {Bread}	0.375	0.75	1.5
{Butter} -> {Milk}	0.375	0.75	1.25
{Bread, Milk} -> {Butter}	0.25	0.666	1.33
{Bread, Butter} -> {Milk}	0.25	0.666	0.888
{Milk, Butter} -> {Bread}	0.25	0.666	1.066

Data and Data Sources

Data Catalogue

A data catalog is essentially an organized inventory of an organization's data assets. It uses metadata (data about data) to make it easier for users to find, understand, and use data.

Features:

- **Metadata Management:** Storing information about data sources, schemas, data lineage, and other relevant details.
- **Data Discovery:** Enabling users to search and find relevant data assets.
- **Data Lineage:** Tracking the origin and flow of data through various systems.
- **Data Governance:** Enforcing policies and controls related to data usage and access.

what-is-a-data-catalog/

Data Catalogue and Data Pipelines

A data catalog provides the "what" and "where" of data, while data pipelines handle the "how" of moving and transforming it.

- A data catalog enhances data pipelines by providing visibility into the data's origin, transformations, and quality.
- Data pipelines populate the data catalog with metadata as data moves through the stages.
- Data catalogs help data engineers and analysts understand the impact of changes to data pipelines.

Data Pipelines

A **data pipeline** is a set of processes that automate the movement, transformation, and processing of data from source to destination.

It ensures data is collected, cleaned, enriched, and stored efficiently for analysis or machine learning.

- [guide-to-data-pipelines/](#)

Key Components of a Data Pipeline

A data pipeline is a series of processes that move and transform data from one or more sources to a destination. Common stages include:

Extraction/Collection: Retrieving data from various sources, such as databases, APIs, files, or streaming platforms.

Ingestion: Bringing the extracted data into a staging area. Batch data ingestion and streaming data ingestion.

Storage: Storing data in a data warehouse, data lake, or a database. In case of ETL this step comes after data preparation.

Key Components of a Data Pipeline

Data Preparation/Transformation:

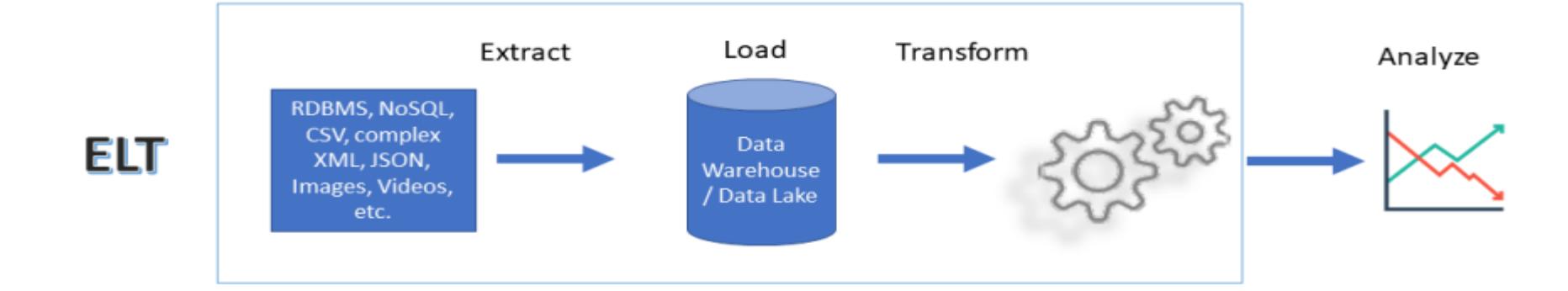
- **Cleaning:** Identifying and correcting errors, inconsistencies, and missing values in the data.
- **Wrangling:** Transforming and structuring the data to make it suitable for analysis or other purposes.
This may involve: Filtering, Aggregating, Joining, Formatting.
- **Exploration and data analysis:** This is where data analysts begin to look at the data, to find patterns, and to understand the data that has been gathered. Querying, reporting, or using data for machine learning.

Versioning & Monitoring : Maintaining a history of data changes, allowing for tracking and rollback if necessary. This is very important for data governance, and for reproducibility of analysis. Managing dependencies, scheduling tasks, and ensuring system reliability.

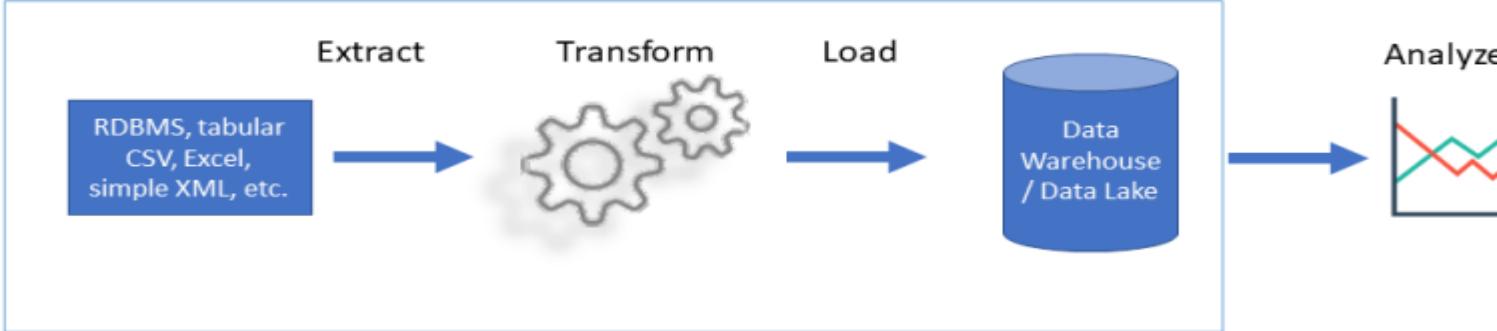
Stages of a Data Pipeline

ELT vs ETL

ELT



ETL



Common Data Pipeline Patterns

Batch Processing Pipeline

- Processes data in chunks at scheduled intervals.
- Suitable for large-scale ETL workloads.
- **Example:** Nightly aggregation of customer transactions for financial reporting.

Technology Stack:

- ◆ Apache Spark, Apache Hadoop, Airflow, AWS Glue

Streaming Data Pipeline

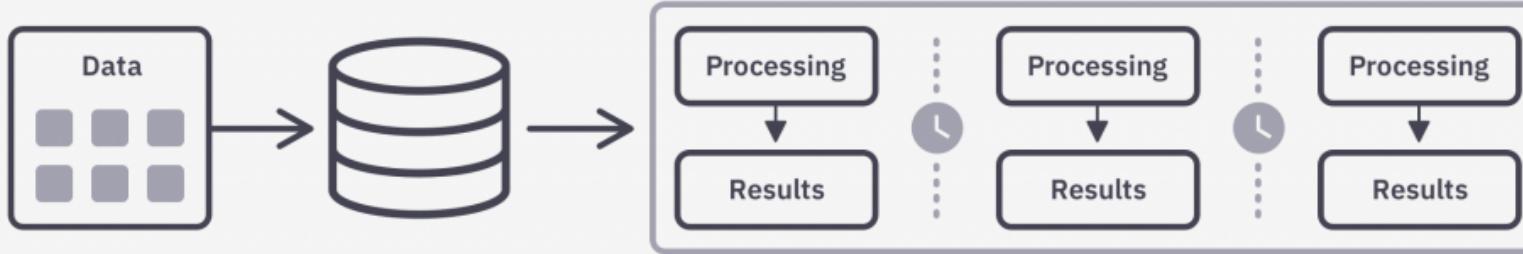
- Processes data in real-time or near real-time.
- Suitable for applications like fraud detection, live analytics, and IoT.
- **Example:** Monitoring website clicks or detecting fraudulent credit card transactions.

Technology Stack:

- ◆ Apache Kafka, Apache Flink, Spark Streaming, AWS Kinesis

Common Data Pipeline Patterns

Batch Processing



Data Stream Processing



Feature	Batch Processing	Stream Processing
Data Processing	Processes a large volume of data at once.	Processes data as it arrives, record by record.
Latency	High latency, as processing happens after data collection.	Low latency, providing near real-time insights.
Throughput	Can handle vast amounts of data at once.	Optimized for real-time but might handle less data volume at a given time.
Use Case	Ideal for historical analysis or large-scale data transformations.	Best for real-time analytics, monitoring, and alerts.
Complexity	Relatively simpler to implement with predefined datasets.	More complex, requires handling continuous streams.
Data Scope	Operates on a finite set of data.	Operates on potentially infinite streams of data.
Error Handling	Errors can be identified and corrected before execution.	Requires real-time handling of errors and failures.
Resource Usage	Resource-intensive during processing, idle otherwise.	Continuous use of resources.
Cost	Cost-effective for large volumes of data.	More expensive due to continuous processing.

Common Data Pipeline Patterns

Lambda Architecture (Hybrid Batch + Stream)

- Combines batch and real-time processing.
- **Example:** A weather app that uses real-time sensor data for short-term forecasts and batch data for long-term trends.

Layers:

- **Batch Layer:** Stores historical data.
- **Speed Layer:** Processes real-time data.
- **Serving Layer:** Merges both for a unified view.

Technology Stack:

- ◆ Apache Kafka, Apache Spark, HDFS, NoSQL Databases

Common Data Pipeline Patterns

Data Lake + Data Warehouse Hybrid

- Stores **raw data** in a **data lake** (e.g., AWS S3, Azure Data Lake).
- Transforms and moves structured data into a **data warehouse** (e.g., Snowflake, Redshift).

Example: An e-commerce company storing all transactions in a data lake but using a warehouse for analytics.

Technology Stack:

- ◆ AWS S3, Azure Data Lake, Snowflake, BigQuery

Best Practices for Data Pipelines

- ✓ **Use a Scalable Architecture** – Design for growing data volume.
- ✓ **Ensure Data Quality** – Use validation and anomaly detection.
- ✓ **Automate Orchestration** – Schedule and monitor pipelines with Apache Airflow.
- ✓ **Optimize Performance** – Use caching, indexing, and parallel processing.
- ✓ **Implement Security & Governance** – Encrypt data, use access controls, and comply with GDPR.

Data Transformation

Need of data transformation:

- To ensure data is consistent and compatible across different systems.
- To improve data quality by cleaning and standardizing it.
- To make data more suitable for specific analytical or modeling needs.

Data transformation includes:

1. Data cleaning:

- Removing or correcting errors, inconsistencies, and duplicates.
- Handling missing values (e.g., imputation).

2. Standardization:

- Formatting data consistently (e.g., date formats, units of measure).
- Normalizing or scaling numerical data.

Data Transformation

3. **Structuring:**
 - Changing the data's organization (e.g., pivoting, aggregating).
 - Converting data types (e.g., string to integer).
4. **Enrichment:** Adding new data or deriving new values from existing data. Joining data from multiple sources.
5. **Filtering:** Removing unwanted data.
6. **Aggregation:** Summarizing data.

Feature Management

Feature Selection

Selecting the most relevant features from a dataset while eliminating redundant or irrelevant ones.

- **Methods:**

- **Filter Methods** (e.g., Correlation, Mutual Information)
- **Wrapper Methods** (e.g., Recursive Feature Elimination)
- **Embedded Methods** (e.g., Lasso Regression)

Feature Engineering

Creating new features from raw data to enhance model learning.

- **Common Techniques:**

- **Polynomial Features** (e.g., $x^2, x^3, x^2 \cdot x^3$)
- **Domain-Specific Transformations** (e.g., Date-Time Feature Extraction)
- **Aggregations and Grouping** (e.g., Mean Purchase Amount per User)

Feature Management

Feature Transformation

Modifying features to meet the assumptions of machine learning algorithms.

- **Techniques:**

- **Scaling** (Standardization, Min-Max Normalization)
- **Encoding** (One-Hot Encoding, Label Encoding)
- **Log Transformations** (for skewed data)

Feature Store & Feature Versioning

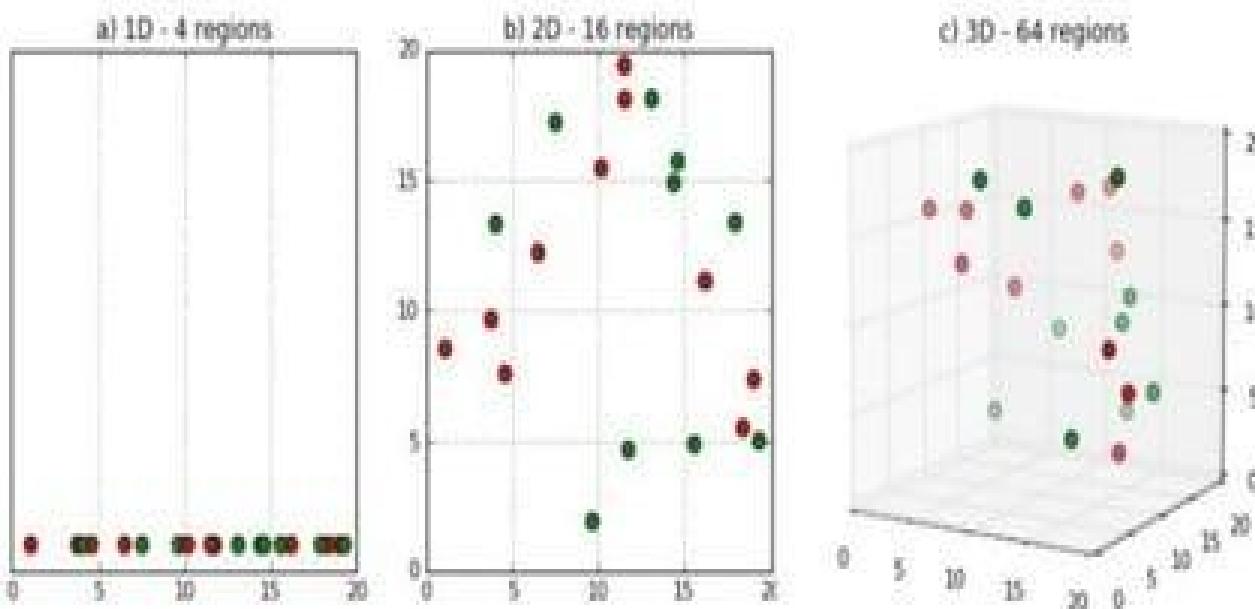
Managing and reusing features efficiently in ML pipelines.

- **Feature Store Tools:** Tecton, Feast, AWS SageMaker Feature Store

- **Feature Versioning:** Tracking feature changes across different ML models.

Principal Component Analysis and LDA

Curse of Dimensionality

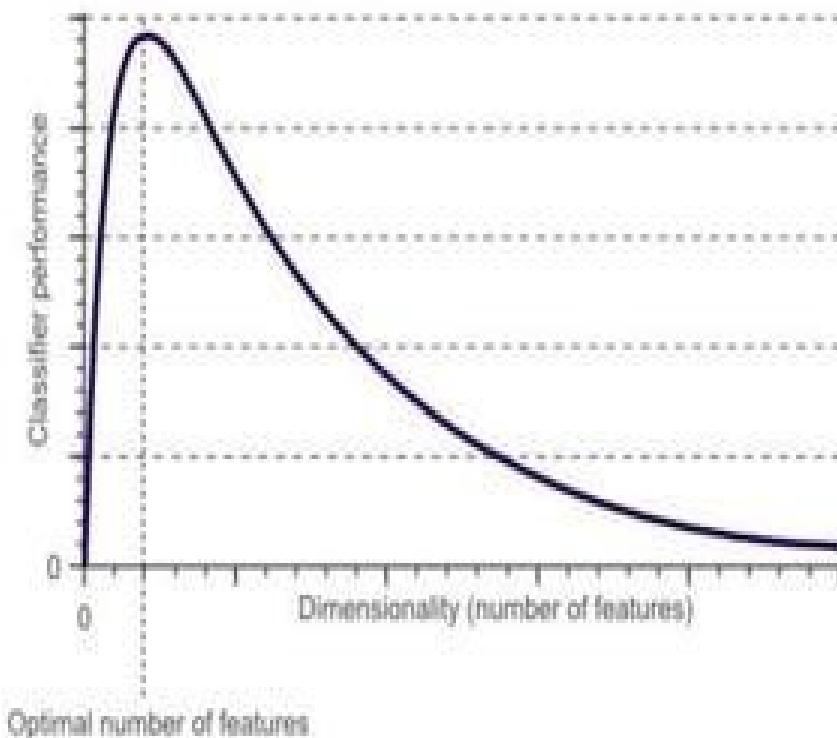


Growth is exponential and everytime we increase the number of dimensions, more data will have to be added to fill the empty spaces.

This exponential growth in data causes high sparsity in the data set and unnecessarily increases storage space and processing time for the particular modelling algorithm. Think of image recognition problem of high resolution images $1280 \times 720 = 921,600$ pixels i.e. 921600 dimensions. OMG. And that's why it's called Curse of Dimensionality.

Curse of Dimensionality

- Refers to the problem caused by the exponential increase in volume associated with adding extra dimensions to a mathematical space.



OVERFITTING

Overfitting occurs when a model starts to memorize the aspects of the training set and in turn loses the ability to generalize



For some algorithms, the number of features controls over-fitting.
E.g. logistic regression

Principal Component Analysis

- A mathematical procedure with ***simple matrix operations*** from ***linear algebra and statistics*** to transform a number of correlated variables into smaller number of uncorrelated variables called principal components.
- Emphasize variation and bring out strong patterns in a dataset.
- To make data easy to ***explore and visualize***.
- Still contains most of the information in the large set.

Principal Component Analysis (contd...)

Figure 1A

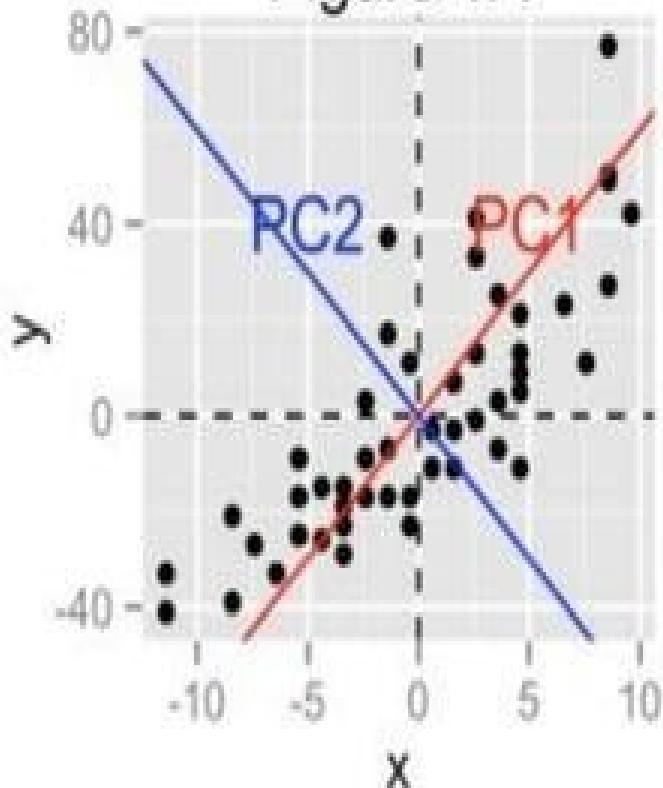


Figure 1B

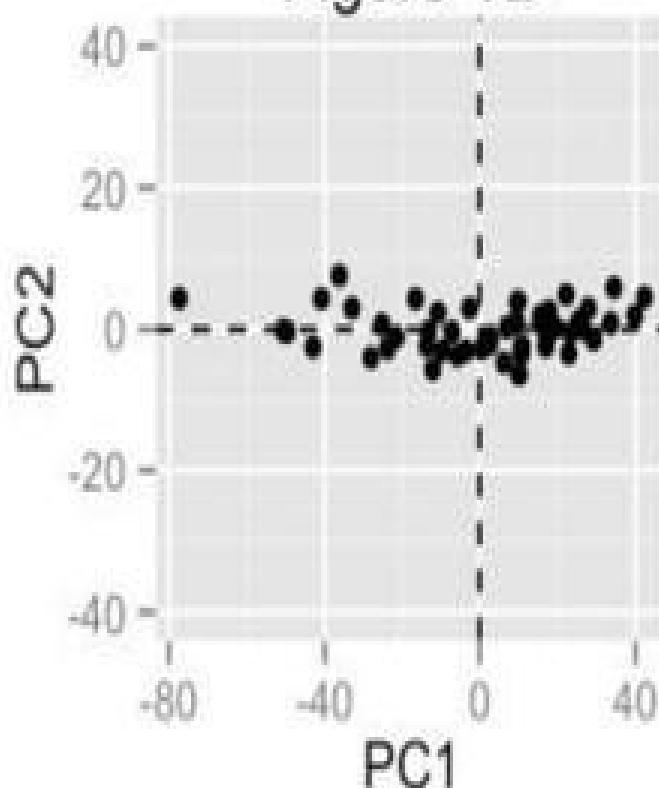


Figure 1A: The data are represented in the X-Y coordinate system

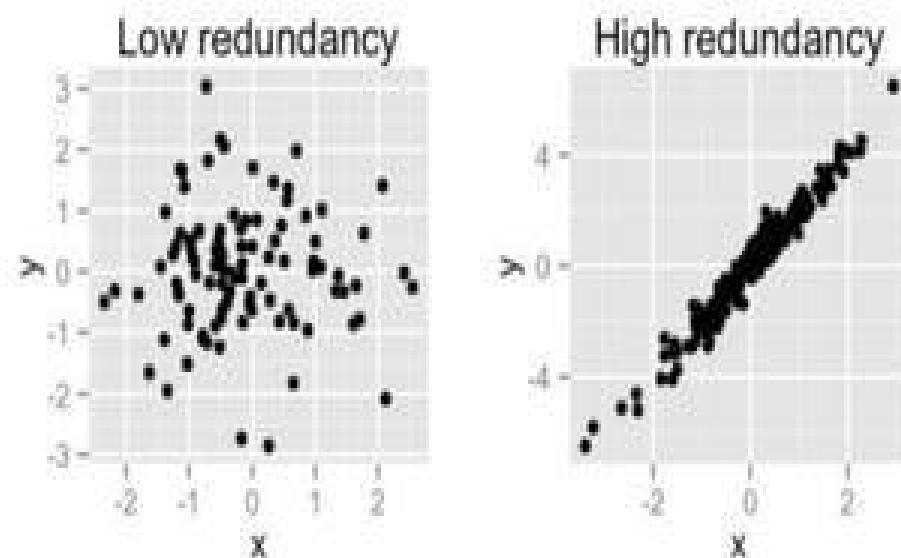
Figure 1B: The **PC1 axis** is the **first principal direction** along which the samples show the largest variation

Goals of PCA

- To *identify hidden pattern* in a data set
- To *reduce the dimensionality* of the data by removing the noise and redundancy in the data
- To identify *correlated variables*

Principal Component Analysis (contd...)

- PCA method is particularly useful when the variables within the data set are highly correlated.
- **Correlation** indicates that there is **redundancy** in the data.
- PCA can be used to reduce the original variables into a smaller number of new variables (= **principal components**) explaining most of the variance in the original variables.



Steps to implement PCA in 2D dataset

Step 1: **Normalize** the data

Step 2: Calculate the **covariance matrix**

Step 3: Calculate the **eigenvalues and eigenvectors**

Step 4: **Choosing** principal components

Step 5: Forming a **feature vector**

Step 6: **Forming Principal Components**

Step 1: Normalization

This is done by subtracting the respective means from the numbers in the respective column. So if we have two dimensions X and Y, all X become x - and all Y become y -.

$$\text{For all } X; \ x = X - \mu_x$$

$$\text{For all } Y; \ y = Y - \mu_y$$

This produces a dataset whose mean is zero.

Step 2: Calculation of correlation

$$\text{Matrix (covariance)} = \begin{bmatrix} \text{var}(x) & \text{var}(x, y) \\ \text{var}(y, x) & \text{var}(y) \end{bmatrix}$$

Covariance Matrix for Iris Dataset

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	0.69	-0.04	1.27	0.52
Sepal.Width	-0.04	0.19	-0.33	-0.12
Petal.Length	1.27	-0.33	3.12	1.30
Petal.Width	0.52	-0.12	1.30	0.58

Calculation of correlation(contd...)

If x and y be two variables with length n ,

The variance of x , variance of y and variance of $x \& y$ is given by following equations.

m_x : mean of x variables

m_y : mean of y variables

$$\sigma_{xx}^2 = \frac{\sum_i (x_i - m_x)(x_i - m_x)}{n - 1}$$

$$\sigma_{yy}^2 = \frac{\sum_i (y_i - m_y)(y_i - m_y)}{n - 1}$$

$$\sigma_{xy}^2 = \frac{\sum_i (x_i - m_x)(y_i - m_y)}{n - 1}$$

Calculation of correlation (contd...)

Correlation is the index to measure how strongly two variable are related to each other. The value of the same ranges from -1 to +1. (i.e. $-1 < r < 1$)

If $r < 0$, variables are ***negatively correlated*** (e.g. x increases when y increases)

If $r > 0$, variables are ***positively correlated*** (e.g. x increases when y decreases)

If $r = 0$, variables has ***no correlation***

Step 3: Calculation of Eigenvalue and Eigenvector

Calculate Eigenvalue and Eigenvector of the covariance matrix using power method:

$$|\lambda - A| = 0$$

Where, I is an identity matrix of same dimension as A
 λ is eigenvalue.

For each value of λ , corresponding eigenvector 'v' is obtained by solving:

$$(\lambda - A)v = 0$$

Step 4: Choosing Component

- Eigenvalues from largest to smallest so that it gives us the components in order of significance.
- If we have a dataset with n variables, then we have the corresponding n eigenvalues and eigenvectors.
- Eigenvector (v) corresponding to largest eigenvalue (λ) is called first principal component.
- To reduce the dimensions, we choose the first p eigenvalues and ignore the rest.

Step 5: Forming a feature vector

Form a feature vector consists of selected eigenvectors

Feature Vector = (eig1, eig2)

Step 5: Forming Principal Components

$$\text{NewData} = \text{FeatureVector}^T \times \text{ScaledData}^T$$

NewData is the Matrix consisting of the principal components,

FeatureVector is the matrix we formed using the eigenvectors we chose to keep,

and

ScaledData is the scaled version of original dataset

Sample Implementation with sklearn

	<u>Output</u>
1 # Principal Component Analysis	
2 from numpy import array	
3 from sklearn.decomposition import PCA	
4 # define a matrix	
5 A = array([[1, 2], [3, 4], [5, 6]])	[[1 2][3 4][5 6]]
6 print(A)	
7 # create the PCA instance	
8 pca = PCA(2)	[[0.70710678 0.70710678]
9 # fit on data	[0.70710678 -0.70710678]]
10 pca.fit(A)	[8.00000000e+00 2.25080839e-33]
11 # access values and vectors ie.covariance	
12 matrix	
13 print(pca.components_)	[[-2.82842712e+00 2.22044605e-16]
14 print(pca.explained_variance_)	[0.00000000e+00 0.00000000e+00]
15 # transform data	
16 B = pca.transform(A)	[2.82842712e+00
17 print(B)	-2.22044605e-16]]

Applications of PCA

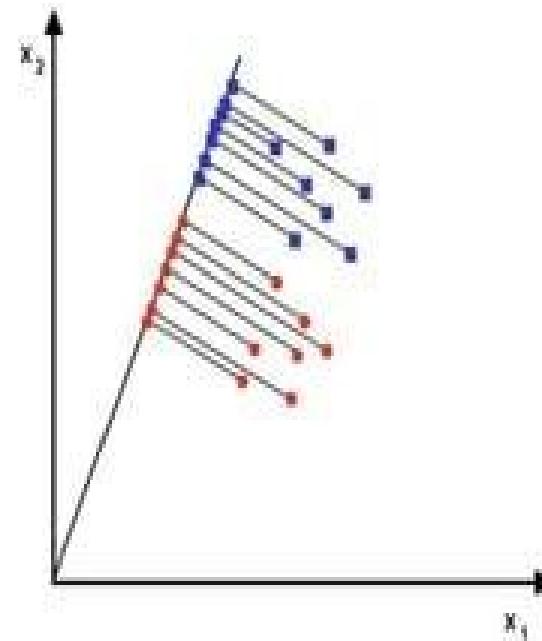
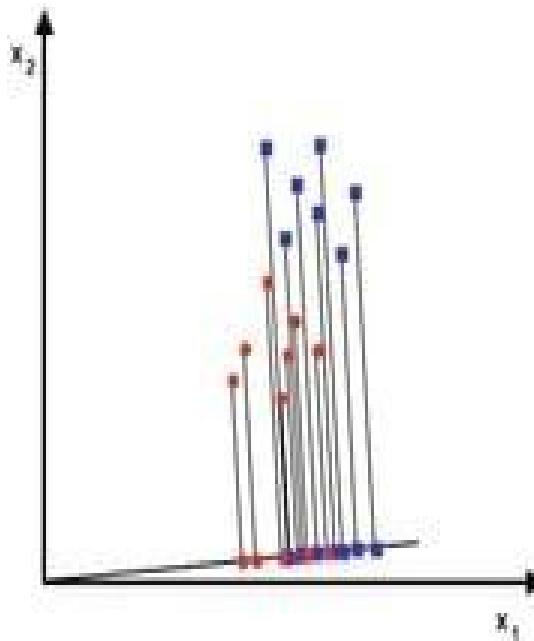
- Commonly used dimensionality reduction technique in domains like *facial recognition, computer vision* and *image compression*.
- Used in *finding patterns in data of high dimension* in the field of *finance, data mining, bioinformatics, psychology*.

Linear Discriminant Analysis (LDA)

- The objective of LDA is to perform dimensionality reduction
- However, we want to preserve as much of the class discriminatory information as possible
- LDA helps you find the boundaries around clusters of classes
- It projects your data points on a line so that your clusters are as separated as possible, with each cluster having a relative (close) distance to a centroid.
- Supervised algorithm as it takes the class label into consideration
- Assume we have a set of D -dimensional samples $x(1), x(2), \dots, x(N)$, N_1 of which belong to class ω_1 , and N_2 to class ω_2
 - We seek to obtain a scalar y by projecting the samples x onto a line $y = w^T x$
 - Of all the possible lines we would like to select the one that maximizes the separability of the scalars

Linear Discriminant Analysis (Cont...)

- It determines a new dimension which is nothing but an axis which should satisfy two criteria:
 - Maximize the distance between the centroid of each class.
 - Minimize the variation (which LDA calls scatter), within each category.



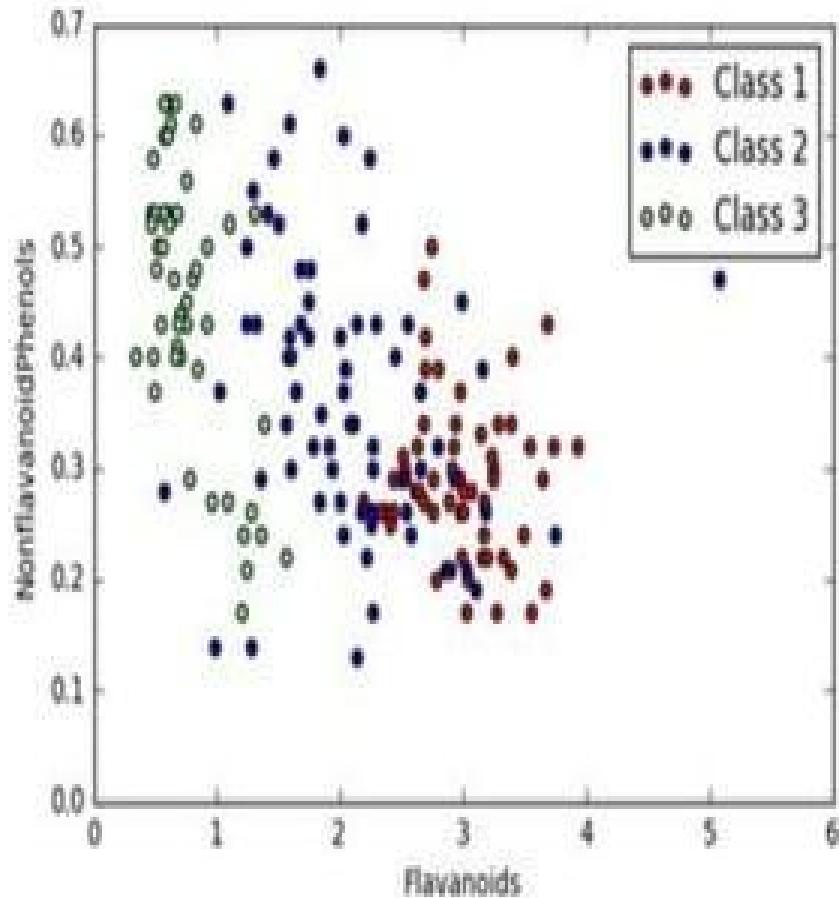


Figure 1.1 : Shows the plotting of data using two features

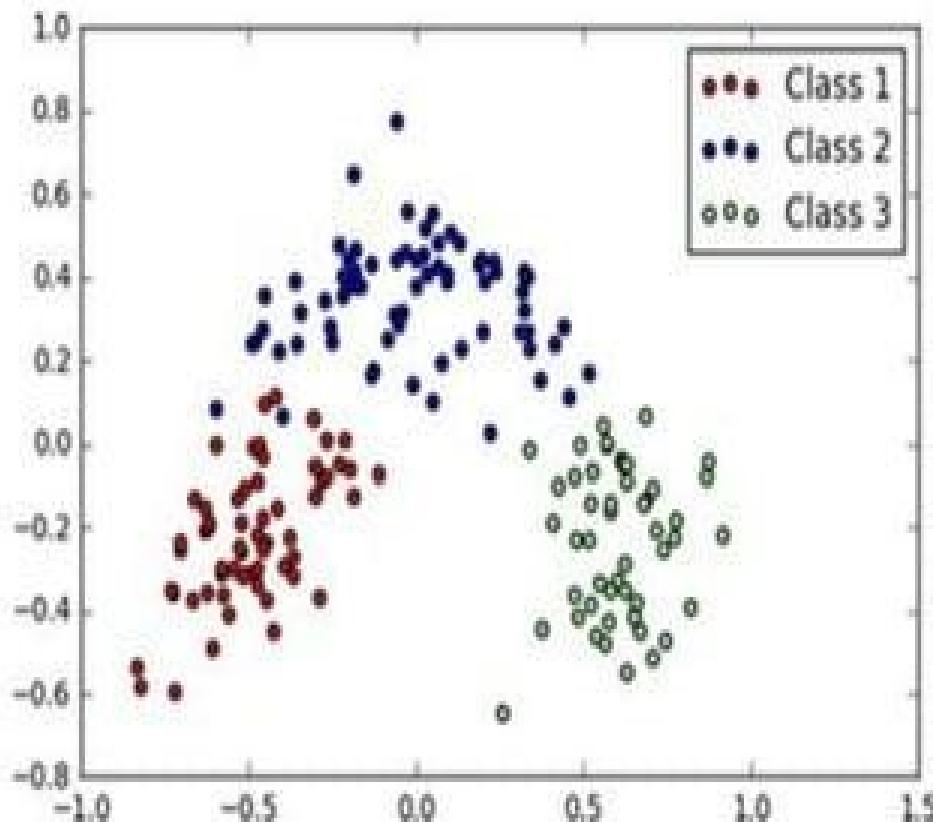
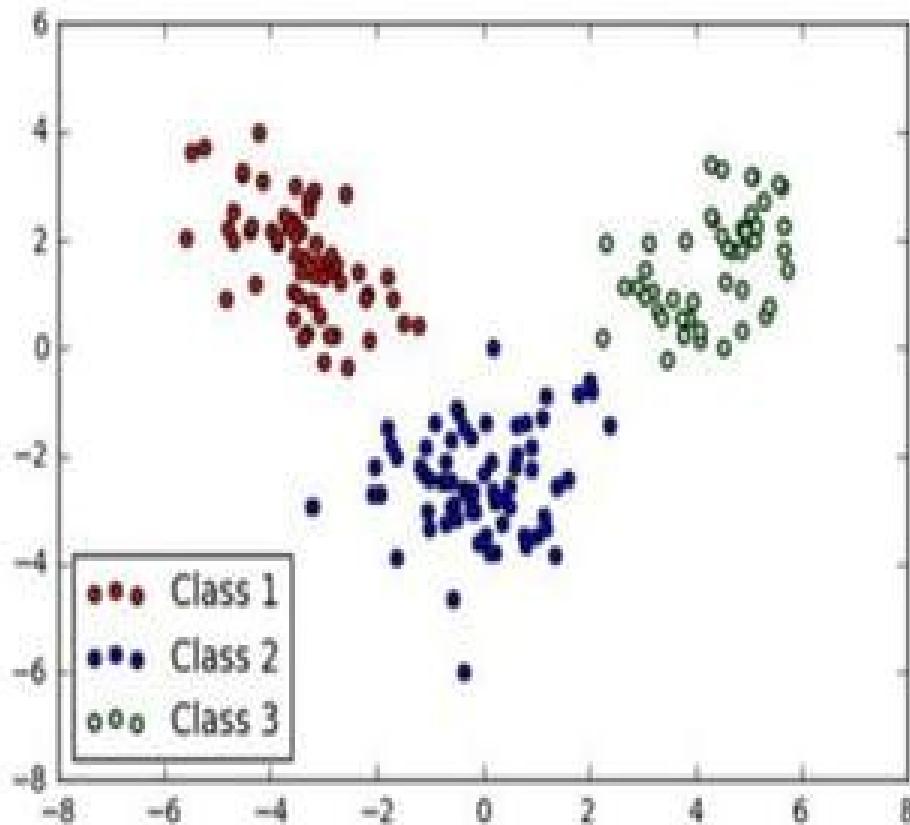
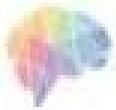


Figure 1.2 : Shows the plotting of data using PCA components



In the graph, you can visualize the whole dataset properly differentiate well among the classes. This is the major difference between PCA and LDA.

Different approaches to LDA

- Class-dependent transformation:
 - Involves maximizing the ratio of between class variance to within class variance
 - The main objective is to maximize this ratio so that adequate class separability is obtained
 - The class-specific type approach involves using two optimizing criteria for transforming the data sets independently.
- Class-independent transformation:
 - This approach involves maximizing the ratio of overall variance to within class variance
 - This approach uses only one optimizing criterion to transform the data sets and hence all data points irrespective of their class identity are transformed using this transform
 - In this type of LDA, each class is considered as a separate class against all other classes

Mathematical Operations

For ease of understanding, this concept is applied to a two-class problem. Each data set has 100 2-D data points.

1. Formulate the data sets and the test sets, which are to be classified in the original space. For ease of understanding let us represent the data sets as a matrix consisting of features in the form given below:

$$\text{Z}_{\text{EIJ}} = \begin{bmatrix} a^{w1} & a^{w2} \\ \dots & \dots \\ \dots & \dots \\ a^{s1} & a^{s2} \\ a^{l1} & a^{l2} \end{bmatrix} \quad \text{Z}_{\text{EIS}} = \begin{bmatrix} p^{w1} & p^{w2} \\ \dots & \dots \\ \dots & \dots \\ p^{s1} & p^{s2} \\ p^{l1} & p^{l2} \end{bmatrix} \quad (1)$$

Mathematical Operations(Cont...)

2. Compute the mean of each data set and mean of entire data set. Let μ_1 and μ_2 be the mean of set 1 and set 2 respectively and μ_3 be mean of entire data, which is obtained by merging set 1 and set 2, is given by Equation 1.

$$\mu_3 = p_1 \times \mu_1 + p_2 \times \mu_2 \quad (2)$$

where p_1 and p_2 are the apriori probabilities of the classes. In the case of this simple two class problem, the probability factor is assumed to be 0.5.

Mathematical Operations(Cont...)

3. In LDA, within-class and between-class scatter are used to formulate criteria for class separability. Within-class scatter is the expected covariance of each of the classes

$$S_w = \sum_j p_j \times (cov_j) \quad (3)$$

Therefore, for the two-class problem,

$$S_w = 0.5 \times cov_1 + 0.5 \times cov_2 \quad (4)$$

Mathematical Operations(Cont...)

Let S_1 and S_2 be the covariance of set 1 and set 2 respectively. Covariance matrix is computed using the following equation:

$$cov_j = (\mathbf{x}_j - \boldsymbol{\mu}_j)(\mathbf{x}_j - \boldsymbol{\mu}_j)^T \quad (5)$$

The between-class scatter is computed using the following equation.

$$S_b = \sum_j (\boldsymbol{\mu}_j - \boldsymbol{\mu}_3) \times (\boldsymbol{\mu}_j - \boldsymbol{\mu}_3)^T \quad (6)$$

Mathematical Operations(Cont...)

The transformations are found as the eigen vector matrix of the different criteria defined in Equations 7 and 8

$$\text{criterion}_j = \text{inv}(\text{cov}_j) \times S_b \quad (7)$$

For the class independent transform, the optimizing criterion is computed as

$$\text{criterion} = \text{inv}(S_w) \times S_b \quad (8)$$

- An eigen vector of a transformation represents a 1-D invariant subspace of the vector space in which the transformation is applied.
- To obtain a non-redundant set of features all eigen vectors corresponding to non-zero eigen values only are considered and the ones corresponding to zero eigen values are neglected.

Mathematical Operations(Cont...)

5. Obtained the transformation matrices, we transform the data sets using the single LDA transform or the class specific transforms which ever the case may be.

For the class dependent LDA,

$$\text{transformed_set}_j = \text{transform}_j^T \times \text{set}_j \quad (9)$$

For the class independent LDA,

$$\text{transformed_set} = \text{transform_spec}^T \times \text{data_set}^T \quad (10)$$

Similarly the test vectors are transformed and are classified using the euclidean distance of the test vectors from each class mean.

Mathematical Operations(Cont...)

- Once the transformations are completed using the LDA transforms, Euclidean distance or RMS distance is used to classify data points.
- Euclidean distance is computed to get the mean of the transformed data set
- The smallest Euclidean distance among the n distances classifies the test vector as belonging to class n.

PCA vs. LDA

- LDA and PCA are linear transformation techniques: LDA is a supervised whereas PCA is unsupervised – PCA ignores class labels.
- PCA performs better in case where number of samples per class is less. Whereas LDA works better with large dataset having multiple classes; class separability is an important factor while reducing dimensionality.
- PCA does more of feature classification and LDA does data classification.
- In PCA, the shape and location of the original data sets changes when transformed to a different space whereas LDA doesn't change the location but only tries to provide more class separability and draw a decision region between the given classes

References

- LDA and PCA for dimensionality reduction
[<https://sebastianraschka.com/faq/docs/lda-vs-pca.html>]
- A. M. Martinez and A. C. Kak, "PCA versus LDA," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228-233, Feb. 2001.
- Principal Component Analysis
[<http://setosa.io/ev/principal-component-analysis/>]

References (contd...)

- **Introduction to Principal Components and Factor Analysis**
[<ftp://statgen.ncsu.edu/pub/thorne/molevoclass/AtchleyOct19.pdf>]
- **Introduction to PCA**
[<https://www.dezyre.com/data-science-in-python-tutorial/principal-component-analysis-tutorial>]
- **Linear Discriminant Analysis - A Brief Analysis**
[<http://www.sthda.com/english/wiki/print.php?id=206>]



THANK YOU...

Feature Engineering techniques
for
text, images, audio, and video.

Features extraction techniques for images

- Histogram of gradients (HOG)
- Maximally stable extremal regions (MSER)
- Scale-invariant feature transform (SIFT)
- Speeded Up Robust Features (SURF)
- Features from accelerated segment test (FAST)
- Local binary pattern (LBP)
- Local phase Quantization (LPQ)
- Edge detection techniques (Canny etc.)

1. Histogram of Oriented Gradients (HOG)

HOG is a feature descriptor that captures gradient orientation and magnitude distributions in localized image regions. It is widely used in object detection, especially pedestrian detection.

How it works:

1. **Gradient Computation:** Compute horizontal (G_x) and vertical (G_y) gradients using Sobel operators.
2. **Gradient Orientation and Magnitude:** Calculate the gradient magnitude and direction:

$$\text{Magnitude} = \sqrt{G_x^2 + G_y^2}$$

$$\text{Orientation} = \tan^{-1} \left(\frac{G_y}{G_x} \right)$$

3. **Spatial Binning:** Divide the image into small cells (e.g., 8×8).
4. **Histogram Creation:** Form histograms of gradient orientations within each cell.
5. **Block Normalization:** Normalize histograms over overlapping blocks to improve invariance to illumination.
6. **Feature Vector Formation:** Concatenate histograms into a single feature vector.

Example Application:

- Used in pedestrian detection in self-driving cars.

2. Maximally Stable Extremal Regions (MSER)

MSER is a region-based feature detector that identifies stable regions under varying illumination and transformations.

How it works:

1. Convert the image to grayscale.
2. Identify connected components at different intensity thresholds.
3. Select regions that remain stable across multiple thresholds.
4. Store these as MSER regions.

Example Application:

- Used in text detection for Optical Character Recognition (OCR).

3. Scale-Invariant Feature Transform (SIFT)

SIFT detects keypoints and describes them using a robust descriptor, making it scale and rotation-invariant.

How it works:

1. **Scale-space Representation:** Apply Gaussian filters at multiple scales.
2. **Keypoint Detection:** Find extrema in the Difference of Gaussian (DoG) space.
3. **Keypoint Localization:** Use Hessian matrix to refine keypoints.
4. **Orientation Assignment:** Assign dominant gradient orientations to make descriptors rotation-invariant.
5. **Feature Descriptor Computation:** Extract histograms of gradients in local patches around keypoints.

Example Application:

- Used in image stitching for panorama creation.

4. Speeded-Up Robust Features (SURF)

SURF is a faster alternative to SIFT and uses integral images for rapid computation.

How it works:

1. Scale-space Representation: Uses Hessian matrix for fast keypoint detection.
2. Orientation Assignment: Uses Haar wavelet responses.
3. Descriptor Generation: Uses wavelet responses to create a 64 or 128-dimensional feature vector.

Example Application:

- Used in object recognition and tracking in real-time applications.

5. Features from Accelerated Segment Test (FAST)

FAST is a high-speed corner detection algorithm.

How it works:

1. Consider a circular region of 16 pixels around a candidate pixel.
2. If a contiguous set of pixels (e.g., 12 out of 16) are brighter or darker than the central pixel, mark it as a corner.
3. Apply non-maximum suppression to remove weak corners.

Example Application:

- Used in real-time applications like mobile augmented reality.

6. Local Binary Patterns (LBP)

LBP is a texture descriptor that encodes pixel neighborhood relationships.

How it works:

1. Divide the image into small regions.
2. For each pixel, compare its intensity with its 8-neighboring pixels.
3. Assign a binary value (1 if the neighbor is greater, 0 otherwise).
4. Convert the binary pattern into a decimal value.
5. Construct a histogram of LBP values.

Example Application:

- Used in face recognition systems.

7. Local Phase Quantization (LPQ)

LPQ is a blur-invariant texture descriptor.

How it works:

1. Apply a Short-Time Fourier Transform (STFT) to small image regions.
2. Quantize the phase information into a binary code.
3. Create a histogram of phase-coded values.

Example Application:

- Used in blur-robust biometric recognition.

8. Canny Edge Detector

Canny is an edge detection technique that finds strong edges in an image.

How it works:

- 1. Noise Reduction:** Apply Gaussian smoothing.
- 2. Gradient Computation:** Compute Sobel gradients.
- 3. Non-Maximum Suppression:** Thin edges by keeping only local maxima.
- 4. Hysteresis Thresholding:** Retain edges using high and low thresholds.

Example Application:

- Used in medical image segmentation.

Summary Table

Method	Type	Key Feature	Application
HOG	Feature Descriptor	Gradient orientation histograms	Pedestrian detection
MSER	Feature Detector	Stable connected regions	OCR text detection
SIFT	Feature Detector & Descriptor	Scale and rotation invariance	Image stitching
SURF	Feature Detector & Descriptor	Fast alternative to SIFT	Object recognition
FAST	Feature Detector	Rapid corner detection	Augmented reality
LBP	Texture Descriptor	Binary texture encoding	Face recognition
LPQ	Texture Descriptor	Blur-invariant texture features	Biometric recognition
Canny	Edge Detector	Multi-stage edge detection	Medical imaging

Textual Features

Number of words

Frequency

Parts of speech

Paragraph

Sentences

1. Number of Words

The number of words in a document helps analyze text length, readability, and complexity.

How it Works:

1. Tokenize the text: Split the text into words.
2. Count the total words, including or excluding stop words (common words like "the," "is," "and").

Example:

Text:

"Natural Language Processing is a branch of AI."

- Word Count (including stop words): 7
- Word Count (excluding stop words like "is", "a", "of"): 5

Applications:

- Used in readability analysis (e.g., Flesch-Kincaid readability score).
- Helps in spam detection (e.g., extremely short messages might be spam).

2. Word Frequency

Word frequency measures how often a word appears in a document or corpus. It is useful for text mining, keyword extraction, and sentiment analysis.

How it Works:

1. Tokenize the text into words.
2. Count the occurrences of each word.
3. Normalize (optional): Convert counts into percentages.

Example:

Text:

"Data science is fun. Data science involves statistics."

- **Word Frequencies:**

- "data" → 2
- "science" → 2
- "is" → 1
- "fun" → 1

- "involves" → 1
- "statistics" → 1

Applications:

- Used in search engines (higher frequency words help in ranking).
- Helps in sentiment analysis (e.g., frequency of "happy" vs. "sad").

4. Paragraph Detection

Paragraph detection involves segmenting text into meaningful sections. A paragraph is a collection of related sentences.

How it Works:

1. Split the text based on line breaks (`\n\n`) or indentation.
2. Count the number of paragraphs.

Example:

Text:

```
pgsql
```

 Copy

 Edit

AI is transforming industries. It improves efficiency and reduces costs.

Machine learning, a subset of AI, enables computers to learn from data.

- Number of Paragraphs: 2

Applications:

- Used in document summarization (each paragraph may discuss a separate topic).
- Helps in document structuring (e.g., AI-based formatting).

5. Sentence Detection

Sentence detection breaks text into individual sentences for grammatical analysis.

How it Works:

1. Split text using punctuation (. ! ?) and spacing.
2. Identify sentence boundaries carefully (e.g., "Dr. Smith is here." is one sentence, not two).

Example:

Text:

"NLP is amazing! It helps machines understand language. What do you think?"

- Sentences:
 1. NLP is amazing!
 2. It helps machines understand language.
 3. What do you think?
- Sentence Count: 3

Applications:

- Used in chatbots and voice assistants (breaking responses into sentences).
- Helps in grammar checking (e.g., Grammarly detects run-on sentences).

Features extraction techniques for textual data

Bag of words

Term frequency-inverse document frequency

Word Embeddings

Feature Extraction Techniques for audio data

Features extraction techniques for audio data

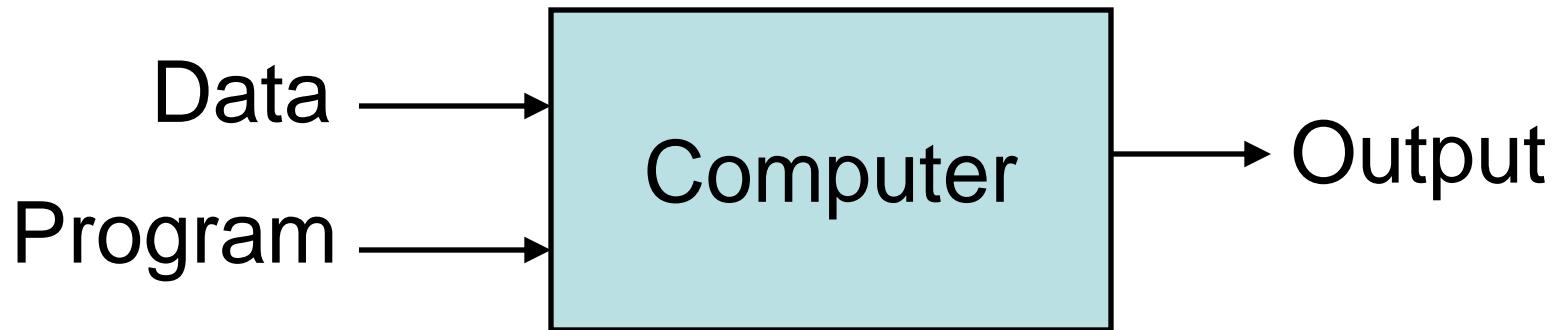
- Mel frequency cepstral coefficients (MFCCs)
- Linear Prediction Coefficient (LPC)
- Linear Prediction Cepstral Coefficients (LPCC)
- Line Spectral Frequencies (LSF)
- Discrete Wavelet Transform (DWT)
- Perceptual Linear Prediction (PLP)

Machine Learning

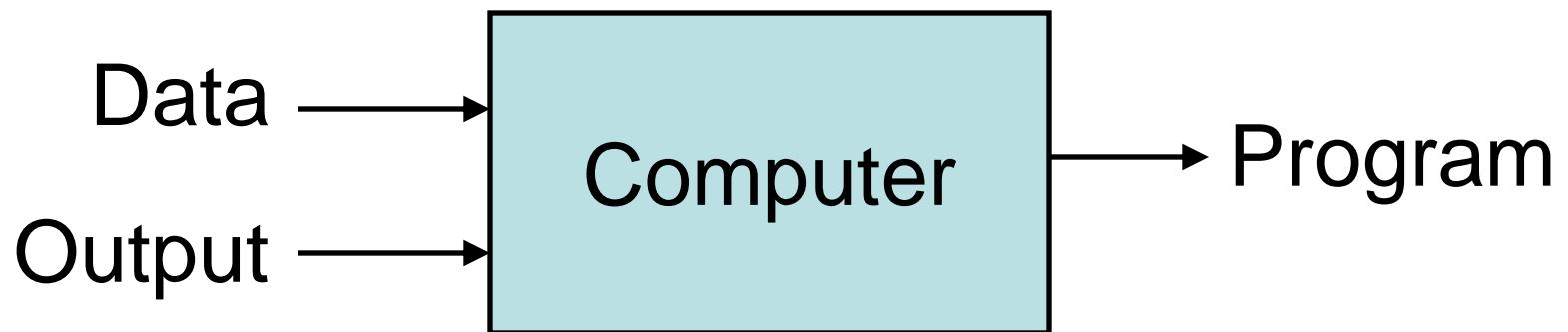
So, What Is Machine Learning?

- Automating automation
- Getting computers to program themselves
- Writing software is the bottleneck
- Let the data do the work instead!

Traditional Programming



Machine Learning



Sample Applications

- Web search
- Computational biology
- Finance
- E-commerce
- Space exploration
- Robotics
- Information extraction
- Social networks
- Debugging
- ...[Your favorite area]

What is Machine Learning?

- Machine Learning
 - Study of algorithms that
 - improve their performance
 - at some task
 - with experience
- Optimize a performance criterion using example data or past experience.
- **Role of Statistics: Inference from a sample**
- **Role of Computer science: Efficient algorithms to**
 - Solve the optimization problem
 - Representing and evaluating the model for inference

Growth of Machine Learning

- Machine learning is preferred approach to
 - Speech recognition, Natural language processing
 - Computer vision
 - Medical outcomes analysis
 - Robot control
 - Computational biology
- This trend is accelerating
 - Improved machine learning algorithms
 - Improved data capture, networking, faster computers
 - Software too complex to write by hand
 - New sensors / IO devices
 - Demand for self-customization to user, environment
 - It turns out to be difficult to extract knowledge from human experts → *failure of expert systems in the 1980's.*

ML in a Nutshell

- Tens of thousands of machine learning algorithms
- Hundreds new every year
- Every machine learning algorithm has three components:
 - **Representation Model**
 - **Evaluation**
 - **Optimization**

Representation

- Decision trees
- Sets of rules / Logic programs
- Instances
- Graphical models (Bayes/Markov nets)
- Neural networks
- Support vector machines
- Model ensembles
- Etc.

Evaluation

- Accuracy
- Precision and recall
- Squared error
- Likelihood
- Posterior probability
- Cost / Utility
- Margin
- Entropy
- K-L divergence
- Etc.

Optimization

- Combinatorial optimization
 - E.g.: Greedy search
- Convex optimization
 - E.g.: Gradient descent
- Constrained optimization
 - E.g.: Linear programming

Types of Learning

- **Association Analysis**
- **Supervised (inductive) learning**
 - Training data includes desired outputs
- **Unsupervised learning**
 - Training data does not include desired outputs
- **Semi-supervised learning**
 - Training data includes a few desired outputs
- **Reinforcement learning**
 - Rewards from sequence of actions

Supervised Learning

- **Given** examples of a function $(X, F(X))$
- **Predict** function $F(X)$ for new examples X
 - Discrete $F(X)$: Classification
 - Continuous $F(X)$: Regression
 - $F(X) = \text{Probability}(X)$: Probability estimation

Supervised Learning: Uses

Example: decision trees tools that create rules

- **Prediction of future cases:** Use the rule to predict the output for future inputs
- **Knowledge extraction:** The rule is easy to understand
- **Compression:** The rule is simpler than the data it explains
- **Outlier detection:** Exceptions that are not covered by the rule, e.g., fraud

Unsupervised Learning

- Learning “what normally happens”
- No output
- Clustering: Grouping similar instances
- Other applications: Summarization, Association Analysis
- Example applications
 - Customer segmentation in CRM
 - Image compression: Color quantization
 - Bioinformatics: Learning motifs

Reinforcement Learning

- Topics:
 - Policies: what actions should an agent take in a particular situation
 - Utility estimation: how good is a state (\rightarrow used by policy)
- No supervised output but delayed reward
- Credit assignment problem (what was responsible for the outcome)
- Applications:
 - Game playing
 - Robot in a maze
 - Multiple agents, partial observability, ...

Supervised vs. unsupervised Learning

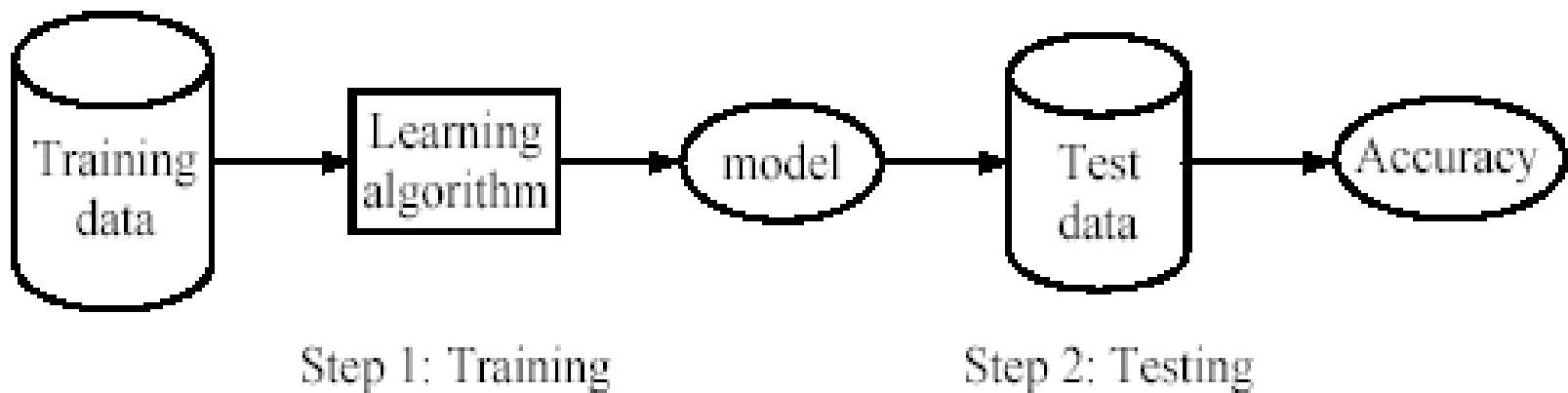
- **Supervised learning:** classification is seen as supervised learning from examples.
 - **Supervision:** The data (observations, measurements, etc.) are labeled with pre-defined classes. It is like that a “teacher” gives the classes (**supervision**).
 - Test data are classified into these classes too.
- **Unsupervised learning (clustering)**
 - Class labels of the data are unknown
 - Given a set of data, the task is to establish the existence of classes or clusters in the data

Supervised learning process: two steps

Learning (training): Learn a model using the **training data**

Testing: Test the model using **unseen test data** to assess the model accuracy

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}},$$



Step 1: Training

Step 2: Testing

Supervised Learning

Supervised vs. Unsupervised Learning

- Supervised learning (classification)
 - Supervision: The training data (observations, measurements, etc.) are accompanied by **labels** indicating the class of the observations
 - New data is classified based on the training set
- Unsupervised learning (clustering)
 - The class labels of training data is unknown
 - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

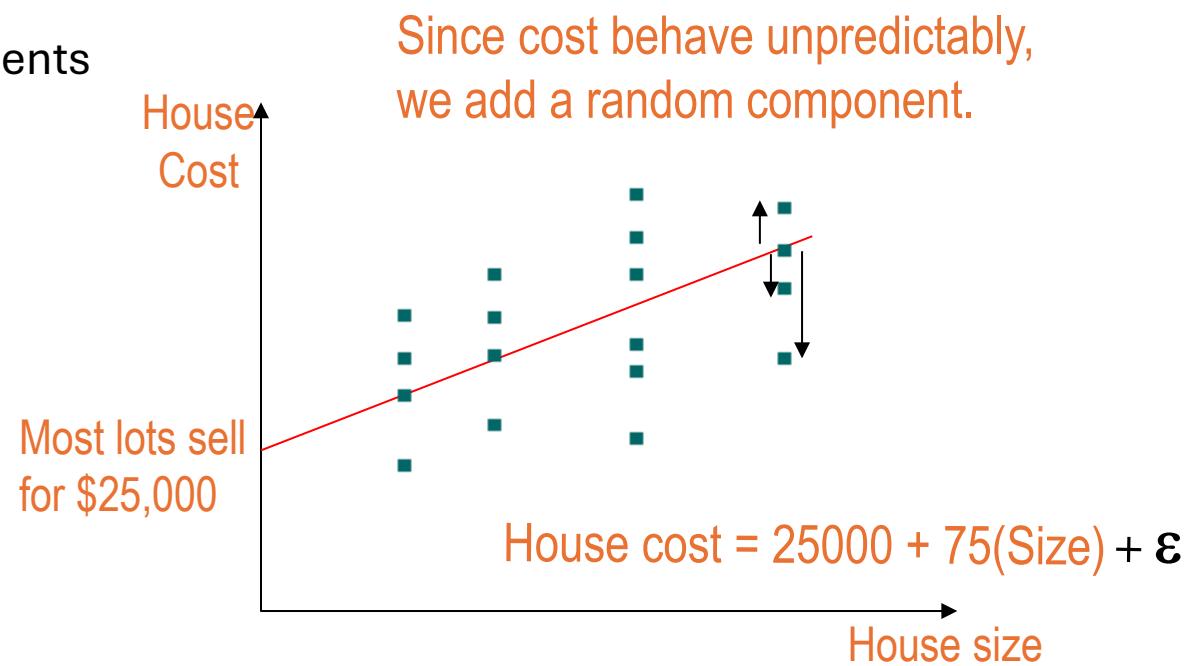
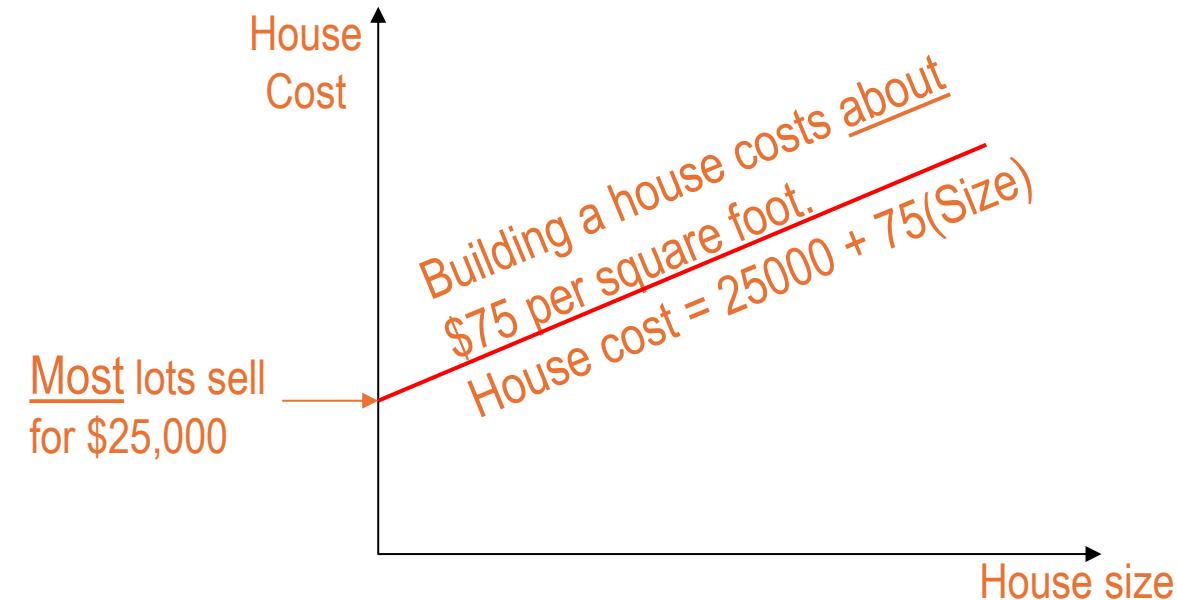
Prediction: Classification vs. Numeric Prediction

- Classification
 - predicts categorical class labels (discrete or nominal)
 - classifies data (constructs a model) based on the training set and the values (**class labels**) in a classifying attribute and uses it in classifying new data
 - E.g. Credit/loan approval, Medical diagnosis: if a tumor is cancerous or benign, Fraud detection: if a transaction is fraudulent or not, etc.
 - **Algorithms:** Decision trees, support vector machines (SVMs), Naive Bayes.
- Numeric Prediction
 - models continuous-valued functions, i.e., predicts unknown or missing values
 - E.g. Predicting house prices, Forecasting stock market values, Estimating temperature, etc.
 - **Algorithms:** Linear regression, polynomial regression, support vector regression (SVR).

Simple linear regression

It is statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

- One variable, denoted x , is regarded as the **predictor, explanatory, or independent** variable.
- The other variable, denoted y , is regarded as the **response, outcome, or dependent** variable.
- We will examine the relationship between quantitative variables x and y via a mathematical equation.
- The model has a deterministic and a statistical components



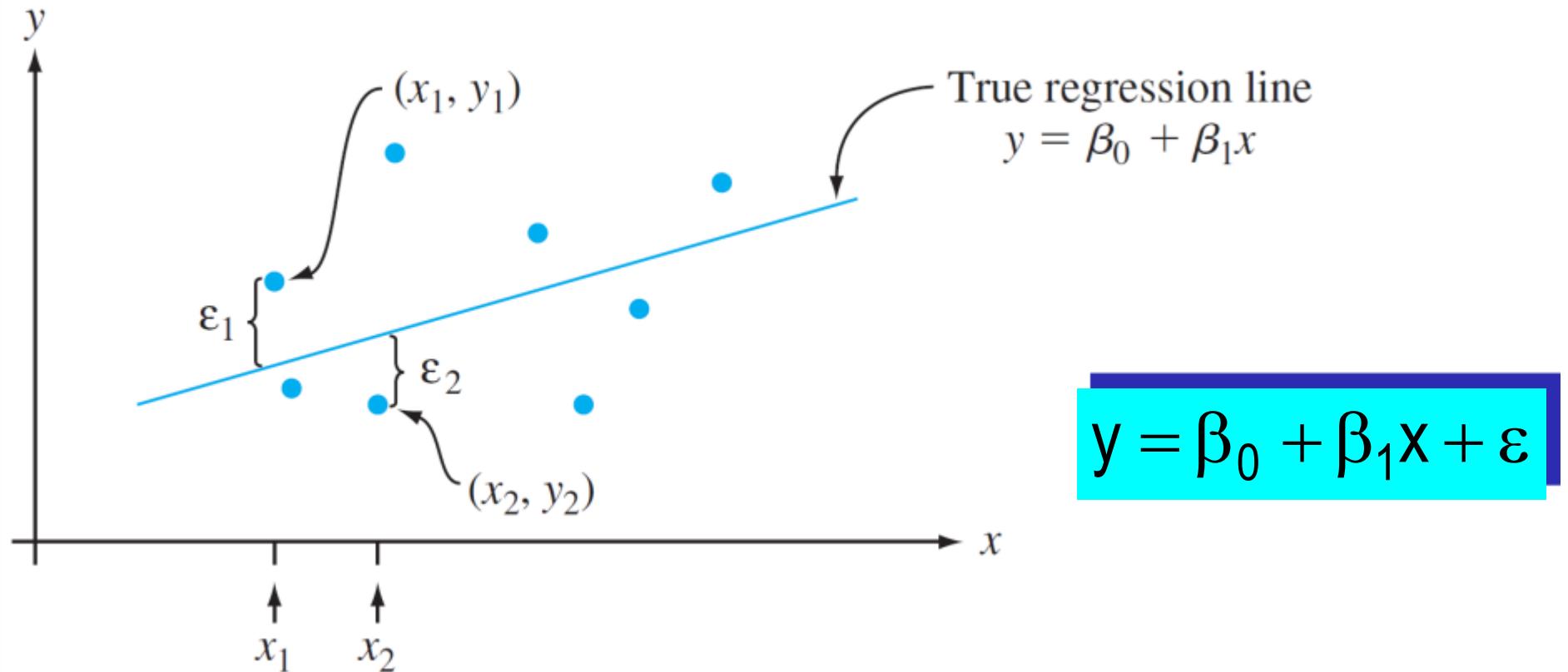
Simple linear regression

- The simplest deterministic mathematical relationship between two variables x and y is a linear relationship: $y = \beta_0 + \beta_1 x$. (True regression line)
- The objective is to develop an equivalent linear probabilistic model.
- If the two (random) variables are probabilistically related, then for a fixed value of x , there is uncertainty in the value of the second variable.
- So, we assume $y = \beta_0 + \beta_1 x + \varepsilon$, where ε is a random variable.

$$b_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad b_0 = \bar{y} - b_1 \bar{x}$$

Simple linear regression

- The points $(x_1, y_1), \dots, (x_n, y_n)$ resulting from n independent observations will then be scattered about the true regression line:



Simple linear regression

Estimating Model parameters:

- The values of β_0 , β_1 and ε will almost never be known to an investigator.
- Instead, sample data consists of n observed pairs $(x_1, y_1), \dots, (x_n, y_n)$, from which the model parameters and the true regression line itself can be estimated.
- Where $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for $i = 1, 2, \dots, n$ and the n deviations $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are independent r.v.'s.
- Aim is to find the **Best Fit Line**: the sum of the squared vertical distances (deviations) from the observed points to that line is as small as it can be.

Simple linear regression

The sum of squared vertical deviations from the points $(x_1, y_1), \dots, (x_n, y_n)$, to the line is then

$$f(b_0, b_1) = \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2$$

The point estimates of β_0 and β_1 , denoted by b_1 and b_0 , are called the least squares estimates – they are those values that minimize using partial derivatives.

$$b_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$SS_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

$$SS_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = (n-1)s_x^2$$

The predicted values are obtained using:

$$\hat{y} = b_0 + b_1 x$$

Simple linear regression

We interpret the fitted value as the value of y that we would predict or expect when using the estimated regression line with $x = x_i$; thus \hat{y}_i is the estimated true mean for that population when $x = x_i$ (based on the data).

The residual $y_i - \hat{y}_i$ is a positive number if the point lies above the line and a negative number if it lies below the line. (x_i, \hat{y}_i)

The residual can be thought of as a measure of deviation and we can summarize the notation in the following way:

$$Y_i - \hat{Y}_i = \hat{\epsilon}_i$$

Simple linear regression

Suppose we have the following data on filtration rate (x) versus moisture content (y):

x	125.3	98.2	201.4	147.3	145.9	124.7	112.2	120.2	161.2	178.9
y	77.9	76.8	81.5	79.8	78.2	78.3	77.5	77.0	80.1	80.2
x	159.5	145.8	75.1	151.4	144.2	125.0	198.8	132.5	159.6	110.7
y	79.9	79.0	76.7	78.2	79.5	78.1	81.5	77.0	79.0	78.6

Relevant summary quantities (*summary statistics*) are

$$\sum x_i = 2817.9, \quad \sum y_i = 1574.8, \quad \sum x_i^2 = 415,949.85,$$

$$\sum x_i y_i = 222,657.88, \quad \text{and} \quad \sum y_i^2 = 124,039.58,$$

From $S_{xx} = 18,921.8295$, $S_{xy} = 776.434$.

Calculation of residuals?

Simple linear regression

x	y	x^2	xy
3	8	9	24
9	6	81	54
5	4	25	20
3	2	9	6
$\Sigma x = 20$	$\Sigma y = 20$	$\Sigma x^2 = 124$	$\Sigma xy = 104$

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$SS_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

$$SS_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = (n-1)s_x^2$$

Using formula,

$$b_1 = \{4*(104) - 20*20\} / \{4*(124) - 20^2\} = 16/96 = 0.166$$

$$b_0 = 20/4 - 0.166*(20/4) = 4.17$$

So, linear regression equation is, $y = b_0 + b_1 x \Rightarrow y = 4.17 + 0.166x$

Simple linear regression

Linear regression, while a powerful tool, has certain limitations that should be considered:

- **Linearity:** Assumes a linear relationship between the dependent and independent variables. If the relationship is non-linear, the model may not accurately capture the underlying pattern.
- **Independence:** Assumes that the errors are independent of each other. If there is autocorrelation in the errors, the model's estimates may be biased and inefficient.
- **Homoscedasticity:** Assumes that the variance of the errors is constant across all levels of the independent variable. If the variance is not constant (heteroscedasticity), the model's estimates may be inefficient.
- **Normality:** Assumes that the errors are normally distributed. If the errors are not normally distributed, the model's inferences may be invalid.
- **Sensitivity to Outliers:** Linear regression can be sensitive to outliers, which can have a significant impact on the model's estimates. Outliers can distort the relationship between the variables and lead to biased results.
- **Limited Flexibility:** Linear regression can only model linear relationships. If the relationship between the variables is complex or non-linear, linear regression may not be able to adequately capture the pattern.

Regression Metrics

Some common regression metrics are

- **Mean Absolute Error (MAE):** $MAE = \frac{1}{n} \sum_{i=1}^n |x_i - y_i|$

- **Mean Squared Error (MSE):** $MSE = \frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2$

- **Root Mean Squared Error (RMSE):** $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2}$

- **R-squared (R^2) Score:** $R^2 = 1 - (\text{SSR} / \text{SST})$

where,

- x_i represents the actual or observed value for the i -th data point.

- y_i represents the predicted value for the i -th data point.

- SSR (Sum of Squared Residuals) and SST (Total Sum of Squares).

$$\begin{aligned} r2_score &= 1 - \frac{\text{total_error_model}}{\text{total_error_baseline}} \\ &= 1 - \frac{\sum_{i=1}^N (\text{predicted}_i - \text{actual}_i)^2}{\sum_{i=1}^N (\text{average_value} - \text{actual}_i)^2} \end{aligned}$$

Regression Metrics

Q. A real estate company is trying to predict the selling price of houses based on their size (in square feet). They trained a regression model and obtained the following predicted prices and actual selling prices for a sample of five houses:

Calculate the MAE, MSE, RMSE, R2 Score.

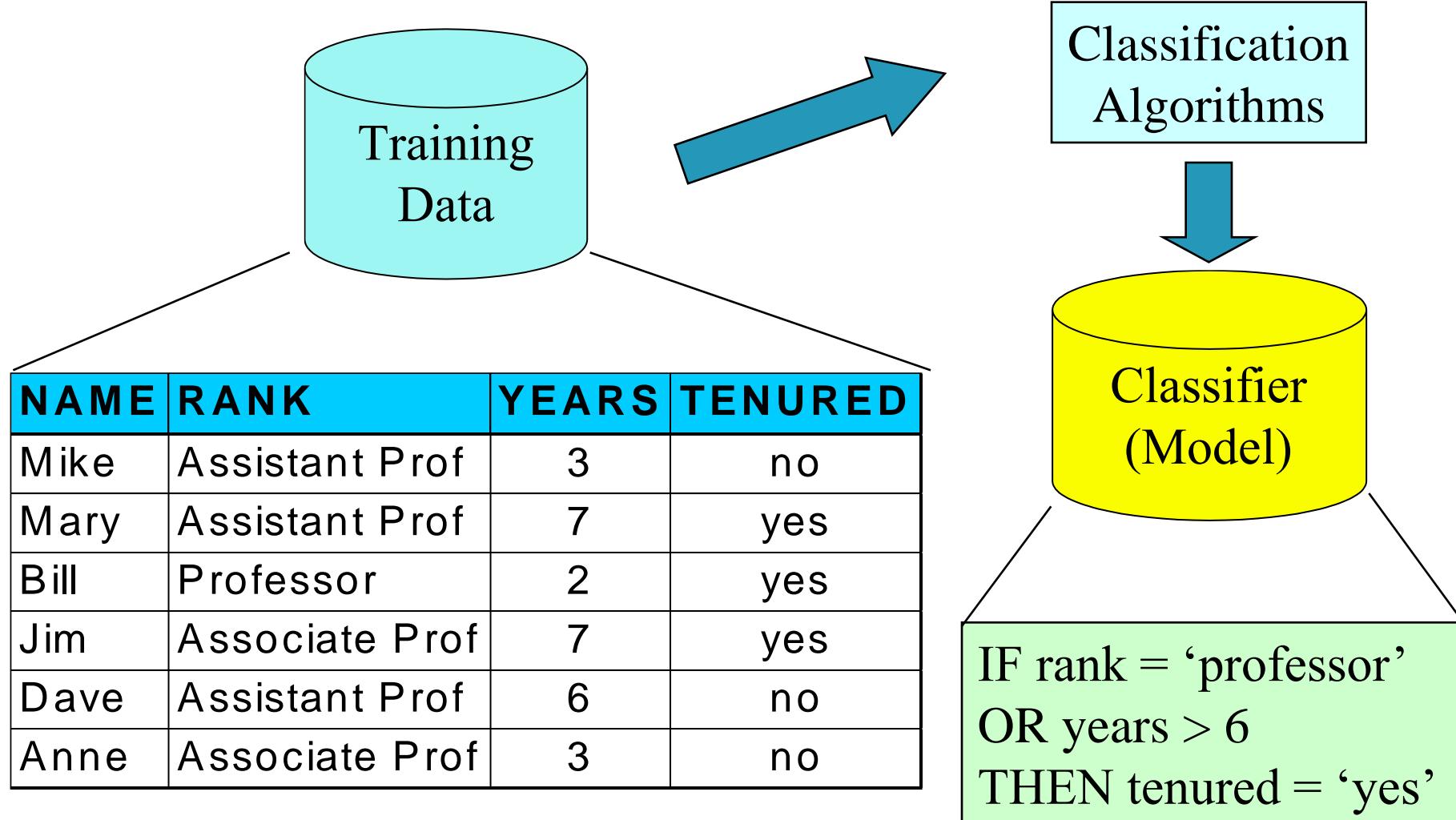
House	Actual Price (in \$1000)	Predicted Price (in \$1000)
1	300	280
2	350	360
3	420	410
4	280	310
5	500	480

Supervised Learning

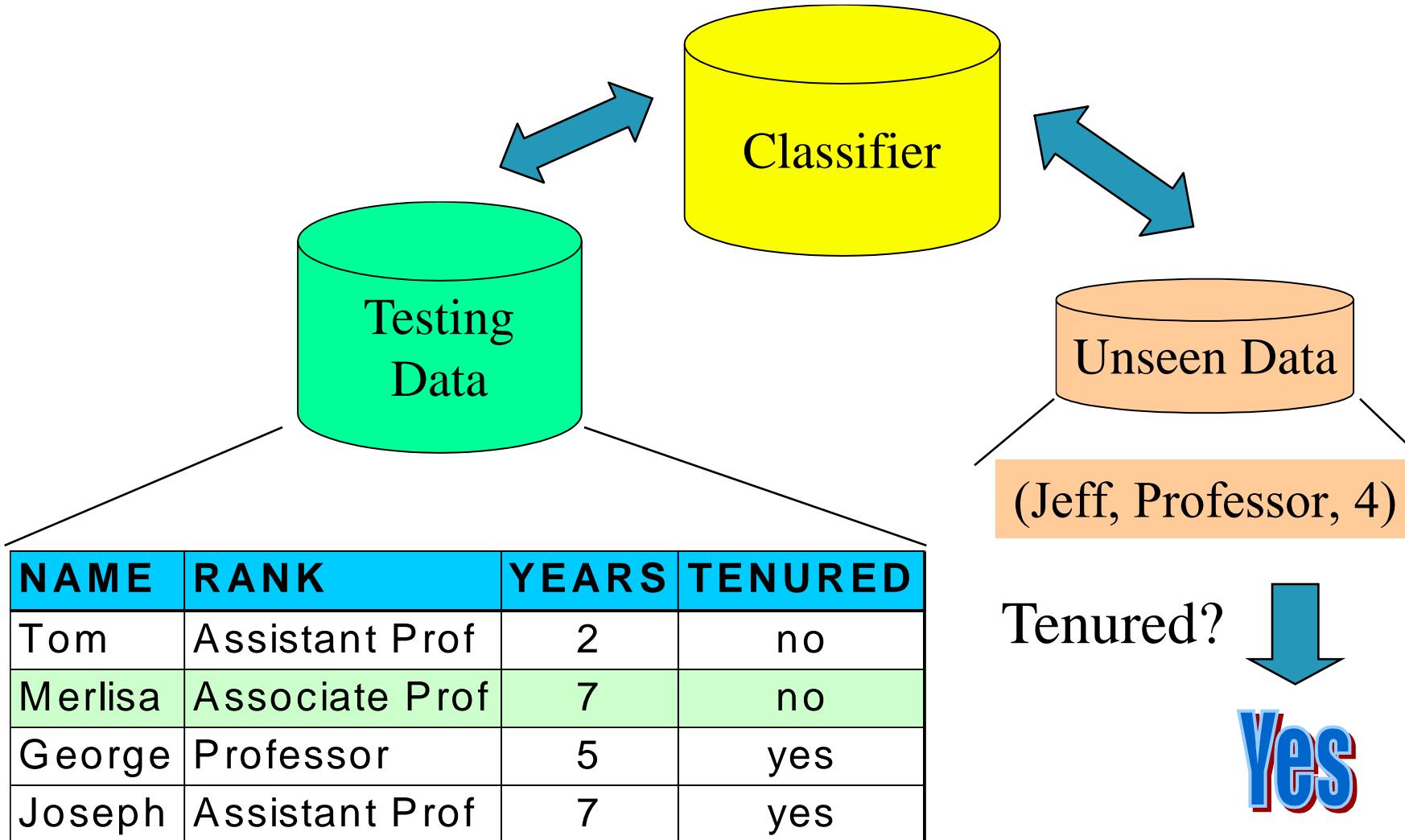
Classification—A Two-Step Process

- Model construction: describing a set of predetermined classes
 - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
 - The set of tuples used for model construction is training set
 - The model is represented as classification rules, decision trees, or mathematical formulae
- Model usage: for classifying future or unknown objects
 - Estimate accuracy of the model
 - The known label of test sample is compared with the classified result from the model
 - Accuracy rate is the percentage of test set samples that are correctly classified by the model
 - Test set is independent of training set (otherwise overfitting)
 - If the accuracy is acceptable, use the model to classify new data
- Note: If *the test set* is used to select models, it is called validation (test) set

Process (1): Model Construction



Process (2): Using the Model in Prediction



Decision Tree

- A decision tree is a tree-like structure where each internal node tests on attribute, each branch corresponds to attribute value and each leaf node represents the final decision or prediction.
- The decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems.

How is Decision Tree formed?

- The process of forming a decision tree involves recursively partitioning the data based on the values of different attributes.
- The algorithm selects the best attribute to split the data at each internal node, based on certain criteria such as information gain or Gini impurity.
- This splitting process continues until a stopping criterion is met, such as reaching a maximum depth or having a minimum number of instances in a leaf node.

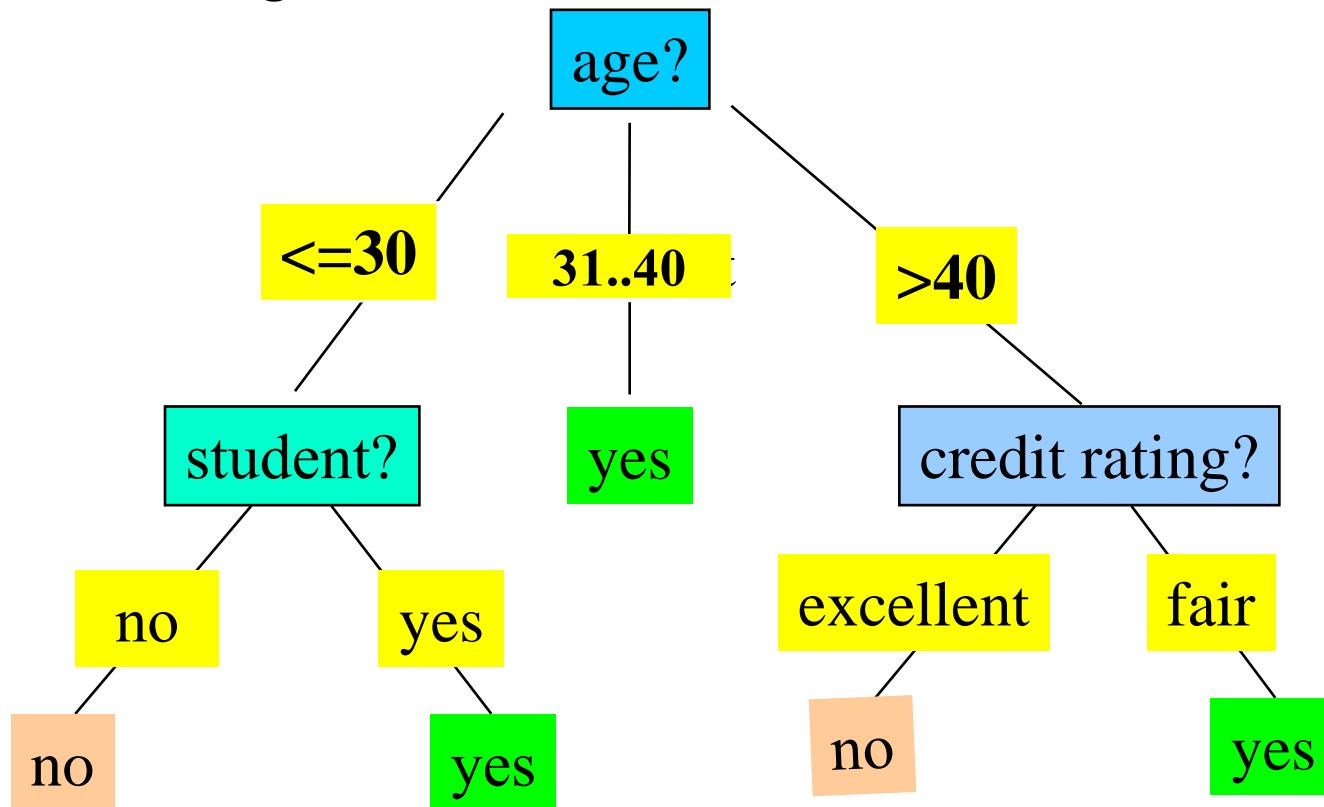
Decision Tree

There are specialized terms associated with decision trees that denote various components and facets of the tree structure and decision-making procedure. :

- **Root Node:** A decision tree's root node, which represents the original choice or feature from which the tree branches, is the highest node.
- **Internal Nodes (Decision Nodes):** Nodes in the tree whose choices are determined by the values of particular attributes. There are branches on these nodes that go to other nodes.
- **Leaf Nodes (Terminal Nodes):** The branches' termini, when choices or forecasts are decided upon. There are no more branches on leaf nodes.
- **Branches (Edges):** Links between nodes that show how decisions are made in response to particular circumstances.
- **Splitting:** The process of dividing a node into two or more sub-nodes based on a decision criterion. It involves selecting a feature and a threshold to create subsets of data.
- **Decision Criterion:** The rule or condition used to determine how the data should be split at a decision node. It involves comparing feature values against a threshold.
- **Pruning:** The process of removing branches or nodes from a decision tree to improve its generalization and prevent overfitting.

Decision Tree Induction: An Example

- ❑ Training data set: Buys_computer
- ❑ The data set follows an example of Quinlan's ID3 (Playing Tennis)
- ❑ Resulting tree:



age	income	student	credit_rating	buys_computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Algorithm for Decision Tree Induction

- Basic algorithm (a greedy algorithm)
 - Tree is constructed in a top-down recursive divide-and-conquer manner
 - At start, all the training examples are at the root
 - Attributes are categorical (if continuous-valued, they are discretized in advance)
 - Examples are partitioned recursively based on selected attributes
 - Test attributes are selected based on a heuristic or statistical measure (e.g., information gain, GINI Index)
- Conditions for stopping partitioning
 - All samples for a given node belong to the same class
 - There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
 - There are no samples left

Attribute Selection Measures

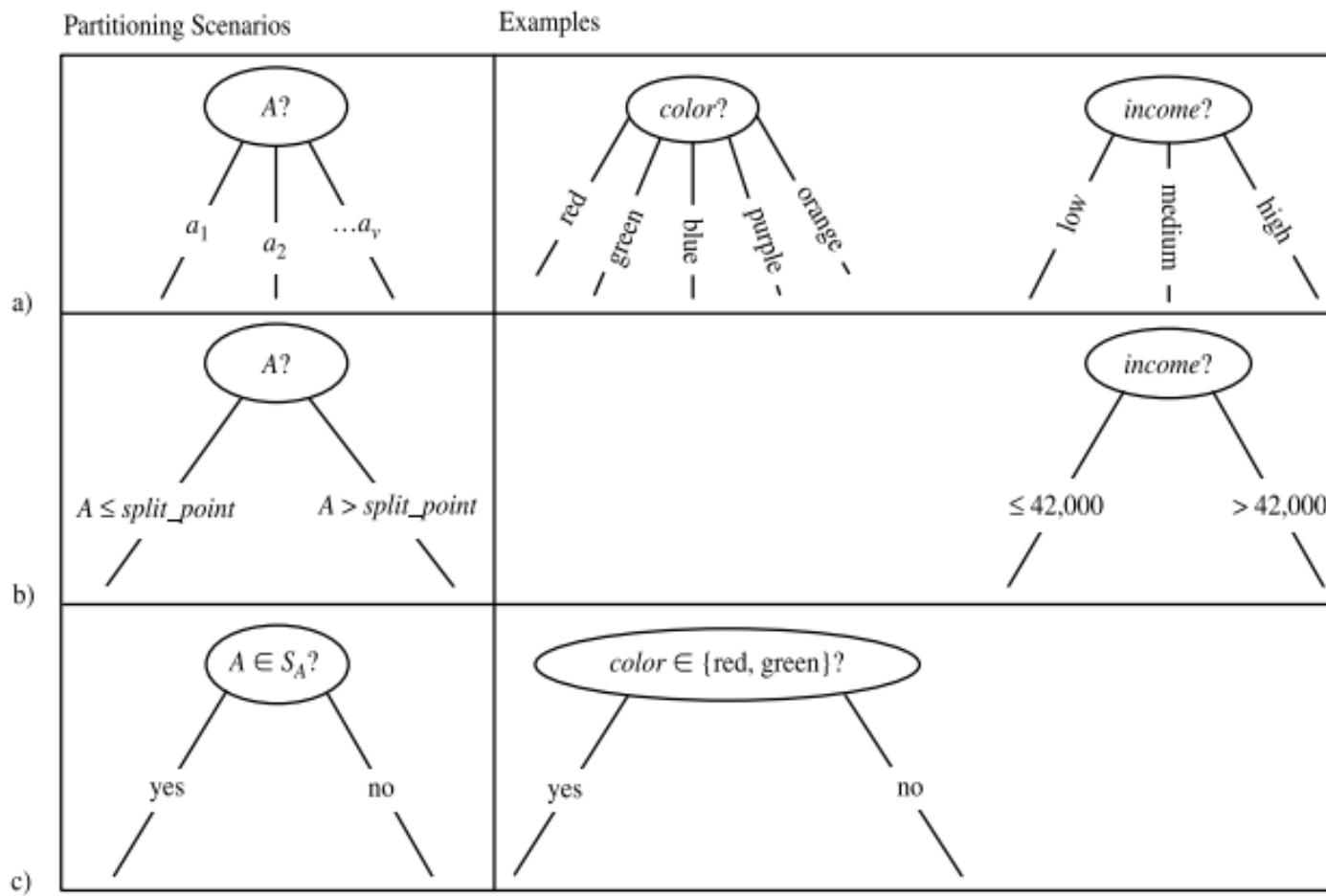
An attribute selection measure is a heuristic for selecting the splitting criterion that “best” separates a given data partition, D, of class-labeled training tuples into individual classes.

Three popular attribute selection measures are :

- information gain,
- gain ratio, and
- gini index.

Let,

- D be the data partition, be a training set of class-labeled tuples.
- Suppose the class label attribute has m distinct values defining m distinct classes, C_i (for $i = 1, \dots, m$).
- Let $C_{i,D}$ be the set of tuples of class C_i in D.
- Let $|D|$ and $|C_{i,D}|$ denote the number of tuples in D and $C_{i,D}$, respectively.



Three possibilities for partitioning tuples based on the splitting criterion, shown with examples. Let A be the splitting attribute. (a) If A is discrete-valued, then one branch is grown for each known value of A. (b) If A is continuous-valued, then two branches are grown, corresponding to $A \leq \text{split_point}$ and $A > \text{split_point}$. (c) If A is discrete-valued and a binary tree must be produced, then the test is of the form $A \in S_A$, where S_A is the splitting subset for A.

Attribute Selection Measures

1. Information Gain:

ID3 uses information gain as its attribute selection measure. When we use a node in a decision tree to partition the training instances into smaller subsets the information/entropy changes.

Information gain is a measure of this change in entropy.

- Expected information/entropy needed to classify a tuple in D is given by $\text{Info}(D) = - \sum_{i=1}^m p_i \log_2(p_i)$, Where, where p_i is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|C_i, D| / |D|$.

■ Now, suppose the tuples in D are partitioned on some attribute A having v distinct values, $\{a_1, a_2, \dots, a_v\}$. D is thus split into v partitions $\{D_1, D_2, \dots, D_v\}$, where D_j contains those tuples in D that have outcome a_j of A.

- We then calculate How much more information would we still need (after the partitioning) in order to arrive at an exact classification? This amount is measured by

$$\text{Info}_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times \text{Info}(D_j).$$

- Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A). That is,
- The attribute A with the highest information gain, ($\text{Gain}(A)$), is chosen as the splitting attribute at node N.

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D).$$

Attribute Selection Measures

Class-labeled training tuples from the *AllElectronics* customer database.

<i>RID</i>	<i>age</i>	<i>income</i>	<i>student</i>	<i>credit_rating</i>	<i>Class: buys_computer</i>
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

$$\text{Info}(D) = - \sum_{i=1}^m p_i \log_2(p_i),$$

$$\text{Info}(D) = - \frac{9}{14} \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right) = 0.940 \text{ bits.}$$

$$\text{Info}_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times \text{Info}(D_j).$$

$$\begin{aligned} \text{Info}_{age}(D) &= \frac{5}{14} \times \left(-\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \right) \\ &\quad + \frac{4}{14} \times \left(-\frac{4}{4} \log_2 \frac{4}{4} - \frac{0}{4} \log_2 \frac{0}{4} \right) \\ &\quad + \frac{5}{14} \times \left(-\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right) \\ &= 0.694 \text{ bits.} \end{aligned}$$

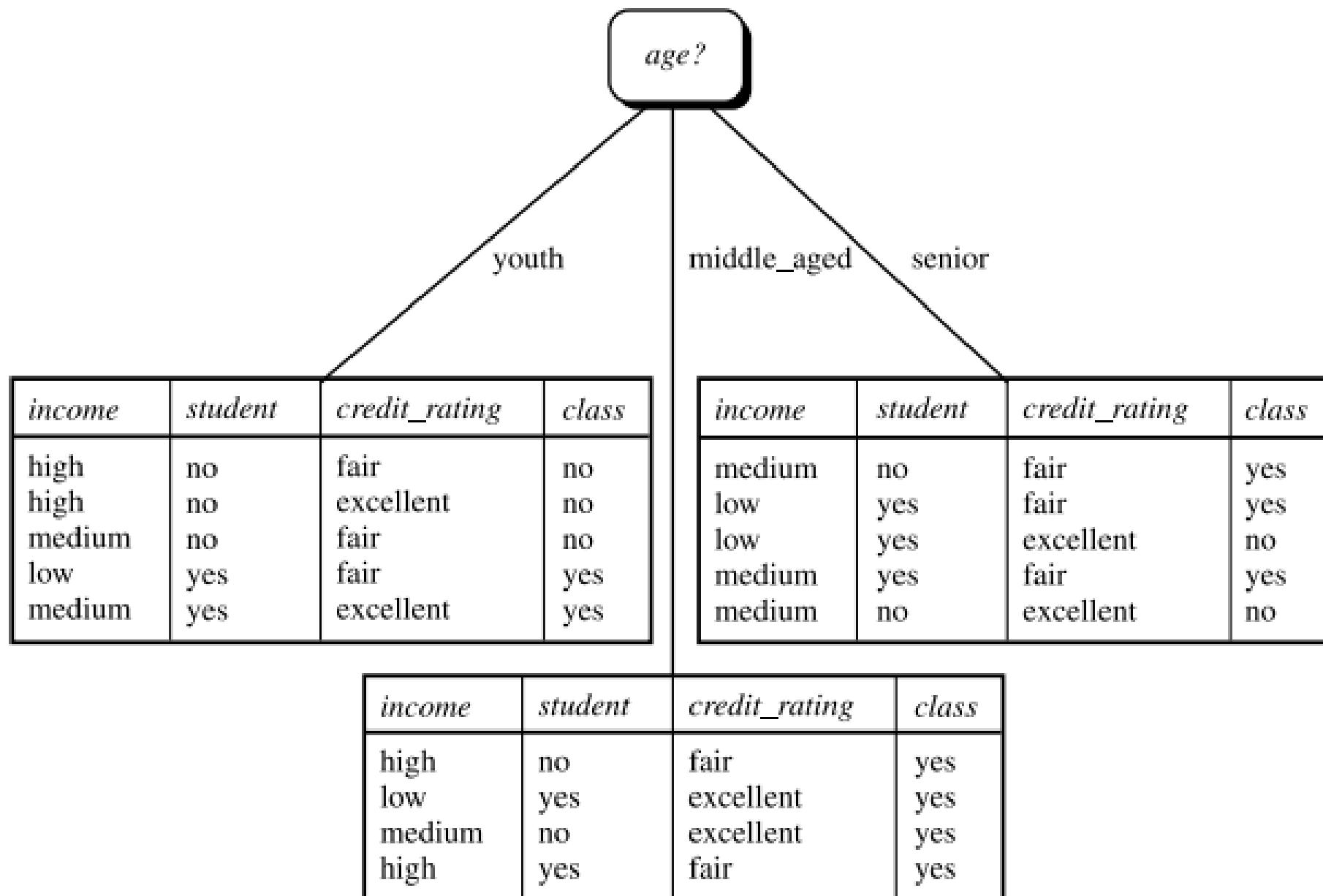
$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D).$$

$$\text{Gain}(\text{age}) = \text{Info}(D) - \text{Info}_{age}(D) = 0.940 - 0.694 = 0.246 \text{ bits.}$$

$$\begin{aligned} \text{Gain}(\text{income}) &= 0.029 \text{ bits}, \text{Gain}(\text{student}) = 0.151 \text{ bits}, \\ \text{Gain}(\text{credit rating}) &= 0.048 \text{ bits.} \end{aligned}$$

Since, **age** has the highest information gain among the attributes, it is selected as the splitting attribute.

Attribute Selection Measures



Attribute Selection Measures

2. Gain Ratio: The information gain measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values.

C4.5, a successor of ID3, uses an extension to information gain known as gain ratio, which attempts to overcome this bias. It applies a kind of normalization to information gain using a “split information” value defined analogously with $\text{Info}(D)$ as

$$\text{SplitInfo}_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{|D|} \right).$$

The gain ratio is defined as

$$\text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}(A)}.$$

Attribute Selection Measures

3. Gini Index

- Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified.
- It means an attribute with a lower Gini index should be preferred.
- The Gini index is used in CART algorithm.
- The Gini value or D is calculated by:

Where, where p_i is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|C_i, D|/|D|$.

$$Gini(D) = 1 - \sum_{i=1}^m p_i^2,$$

- The Gini index considers a binary split for each attribute. To determine the best binary split on A, we examine all of the possible subsets that can be formed using known values of A.
- If A has v possible values, then there are 2^v possible subsets.
 - For example, if income has three possible values, namely {low, medium, high}, then the possible subsets are {low, medium, high}, {low, medium}, {low, high}, {medium, high}, {low}, {medium}, {high}, and {}.
 - We exclude the power set, {low, medium, high}, and the empty set from consideration since, conceptually, they do not represent a split.
 - Therefore, there are $2^v - 2$ possible ways to form two partitions of the data, D, based on a binary split on A.

Attribute Selection Measures

3. Gini Index

The Gini value or D is calculated by:

Where, where p_i is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|C_i|/|D|$.

$$Gini(D) = 1 - \sum_{i=1}^m p_i^2,$$

- The Gini index considers a binary split for each attribute. To determine the best binary split on A, we examine all of the possible subsets that can be formed using known values of A.
- When considering a binary split, we compute a weighted sum of the impurity of each resulting partition. For example, if a binary split on A partitions D into D₁ and D₂, the gini index of D given that partitioning is

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2).$$

The reduction in impurity that would be incurred by a binary split on a discrete- or continuous-valued attribute A is $\Delta Gini(A) = Gini(D) - Gini_A(D)$.

The attribute that maximizes the reduction in impurity (or, equivalently, has the minimum Gini index) is selected as the splitting attribute.

Attribute Selection Measures

Example of Gini index

Let D be the training data that has 9 tuples belonging to the class buys_computer = yes and the remaining 5 tuples belong to the class buys_computer = no.

$$Gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459.$$

Let's start with the attribute income, and its subset {low, medium}. The Gini index value computed based on this partitioning is

$$\begin{aligned} & Gini_{income \in \{low, medium\}}(D) \\ &= \frac{10}{14} Gini(D_1) + \frac{4}{14} Gini(D_2) \\ &= \frac{10}{14} \left(1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2\right) + \frac{4}{14} \left(1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2\right) \\ &= 0.450 \\ &= Gini_{income \in \{high\}}(D). \end{aligned}$$

Similarly, the Gini index values for splits on the remaining subsets are: 0.315 (for the subsets {low, high} and {medium}) and 0.300 (for the subsets {medium, high} and {low}). Therefore, the best binary split for attribute income is on {medium, high} (or {low}) because it minimizes the gini index.

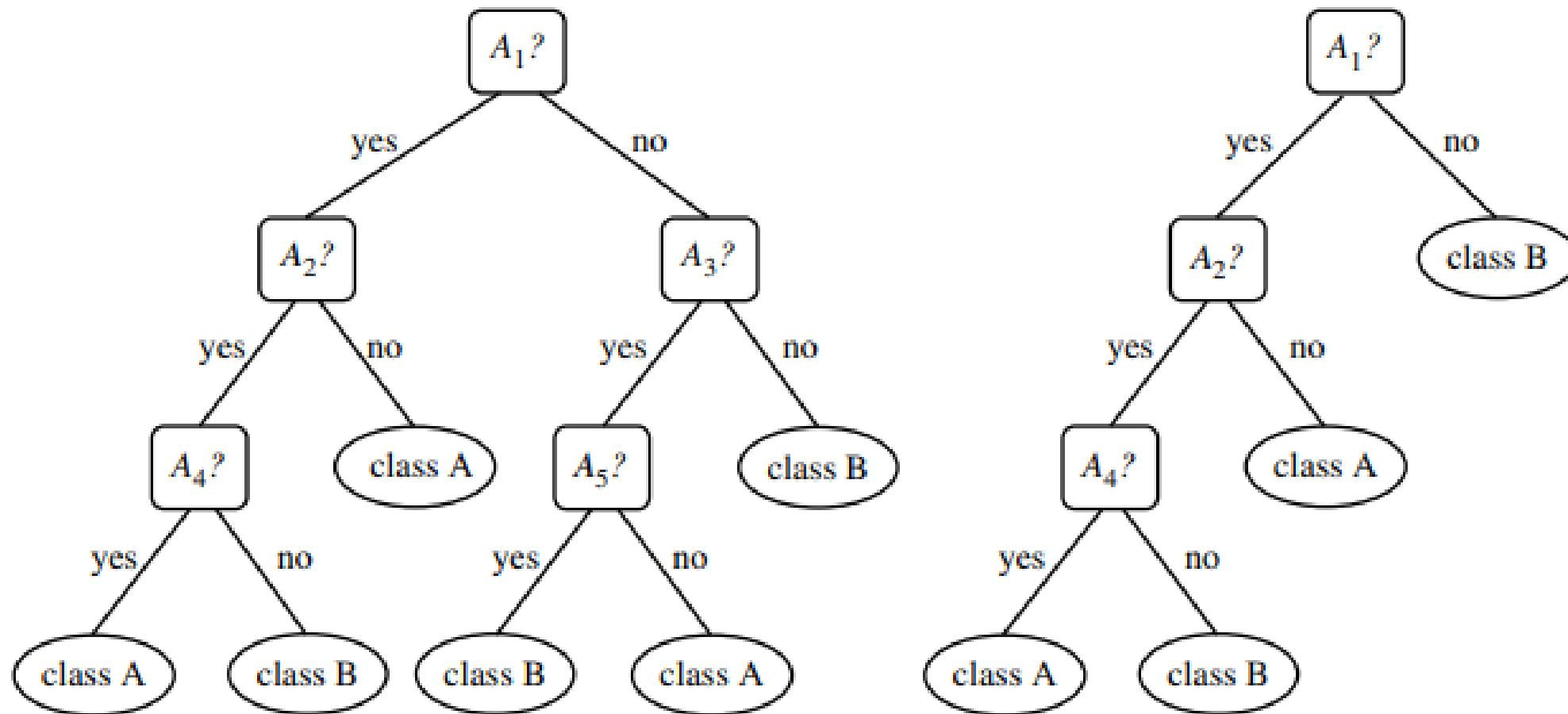
Evaluating the attribute, we obtain {youth, senior} (or {middle aged}) as the best split for age with a Gini index of 0.375; the attributes {student} and {credit rating} are both binary, with Gini index values of 0.367 and 0.429, respectively.

Overfitting and Tree Pruning

- Overfitting: An induced tree may overfit the training data
 - Too many branches, some may reflect anomalies due to noise or outliers
 - Poor accuracy for unseen samples
- Two approaches to avoid overfitting
 - Prepruning: *Halt tree construction early* -do not split a node if this would result in the goodness measure falling below a threshold
 - Difficult to choose an appropriate threshold
 - Postpruning: *Remove branches* from a “fully grown” tree—get a sequence of progressively pruned trees
 - Use a set of data different from the training data to decide which is the “best pruned tree”

Overfitting and Tree Pruning

- Example of pruning



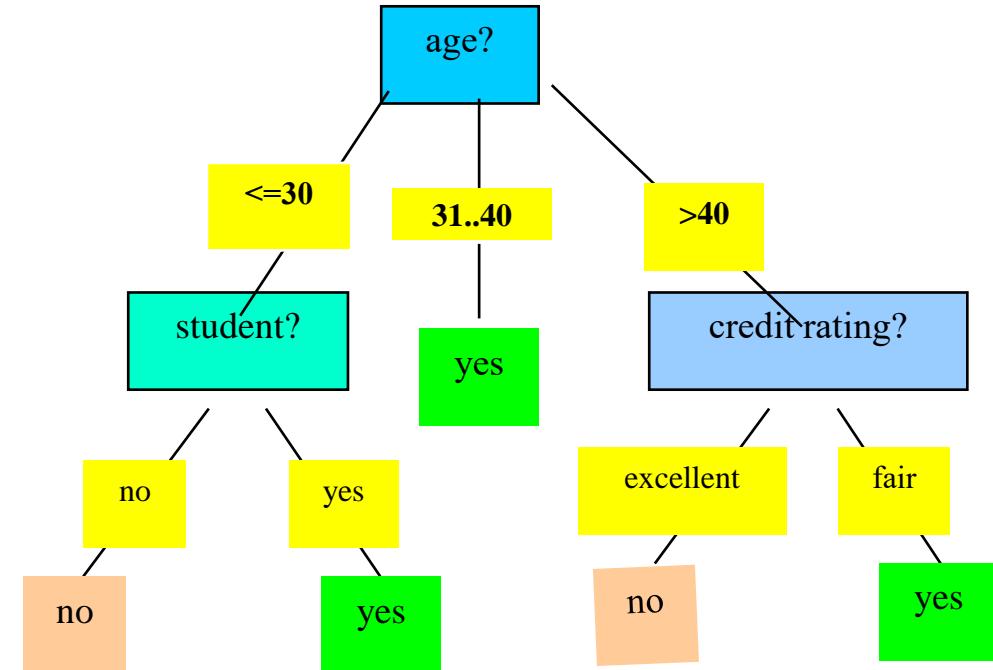
Practice question

[decision-trees-for-classification](#)

Day	Outlook	Temperature	Humidity	Wind	Play Golf
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Rule Extraction from a Decision Tree

- Rules are *easier to understand* than large trees
- One rule is created *for each path* from the root to a leaf
- Each attribute-value pair along a path forms a conjunction: the leaf holds the class prediction
- Rules are mutually exclusive and exhaustive
 - Example: Rule extraction from our *buys_computer* decision-tree



IF age = young AND student = no	THEN buys_computer = no
IF age = young AND student = yes	THEN buys_computer = yes
IF age = mid-age	THEN buys_computer = yes
IF age = old AND credit_rating = excellent	THEN buys_computer = no
IF age = old AND credit_rating = fair	THEN buys_computer = yes

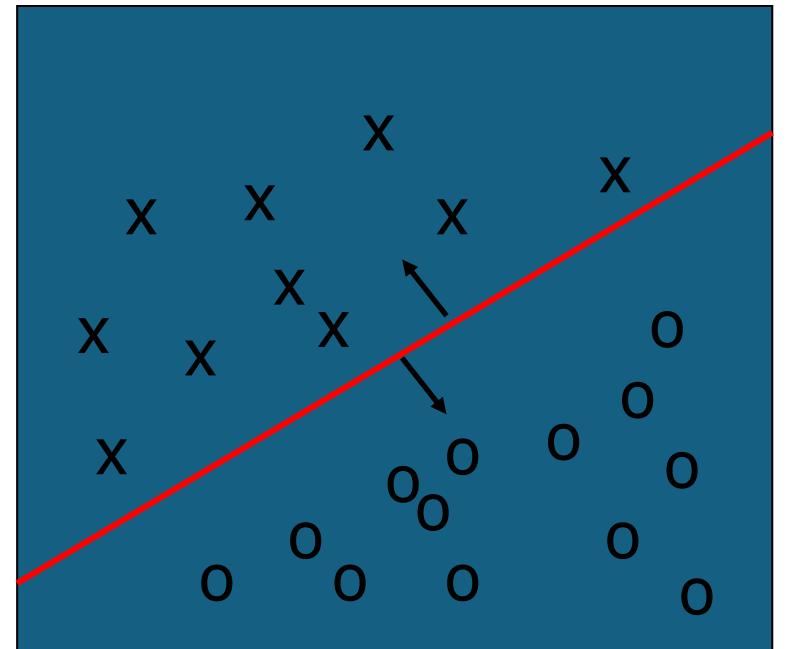
Supervised Learning

Classification—A Two-Step Process

- Model construction: describing a set of predetermined classes
 - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
 - The set of tuples used for model construction is training set
 - The model is represented as classification rules, decision trees, or mathematical formulae
- Model usage: for classifying future or unknown objects
 - Estimate accuracy of the model
 - The known label of test sample is compared with the classified result from the model
 - Accuracy rate is the percentage of test set samples that are correctly classified by the model
 - Test set is independent of training set (otherwise overfitting)
 - If the accuracy is acceptable, use the model to classify new data
- Note: If *the test set* is used to select models, it is called validation (test) set

Classification: A Mathematical Mapping

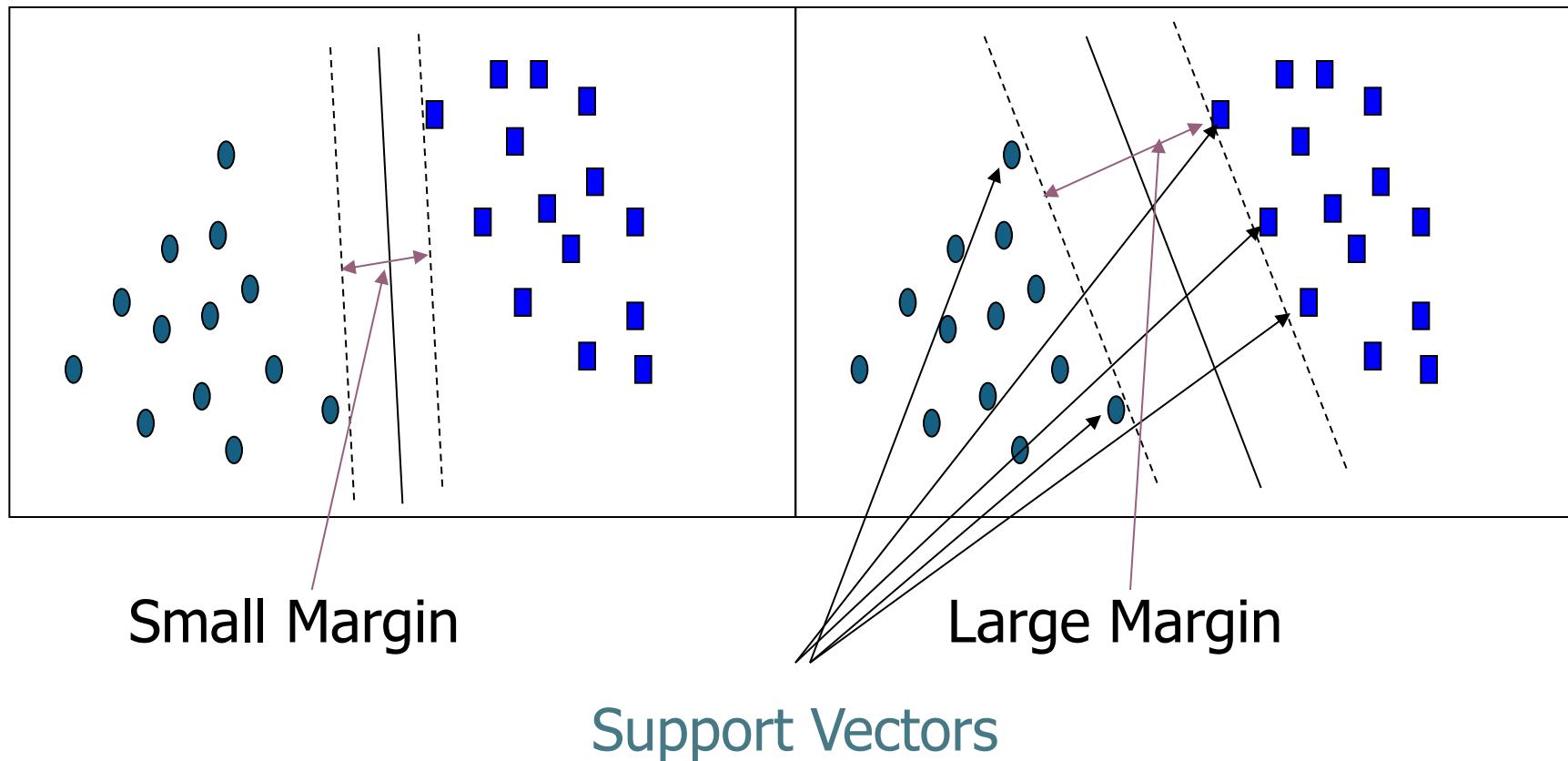
- Classification: predicts categorical class labels
 - E.g., Personal homepage classification
 - $x_i = (x_1, x_2, x_3, \dots)$, $y_i = +1$ or -1
 - x_1 : # of word “homepage”
 - x_2 : # of word “welcome”
 - Mathematically, $x \in X = \mathbb{R}^n$, $y \in Y = \{+1, -1\}$,
 - We want to derive a function $f: X \rightarrow Y$
 - Linear Classification
 - Binary Classification problem
 - Data above the red line belongs to class ‘x’
 - Data below red line belongs to class ‘o’
 - Examples: SVM, Perceptron, Probabilistic Classifiers



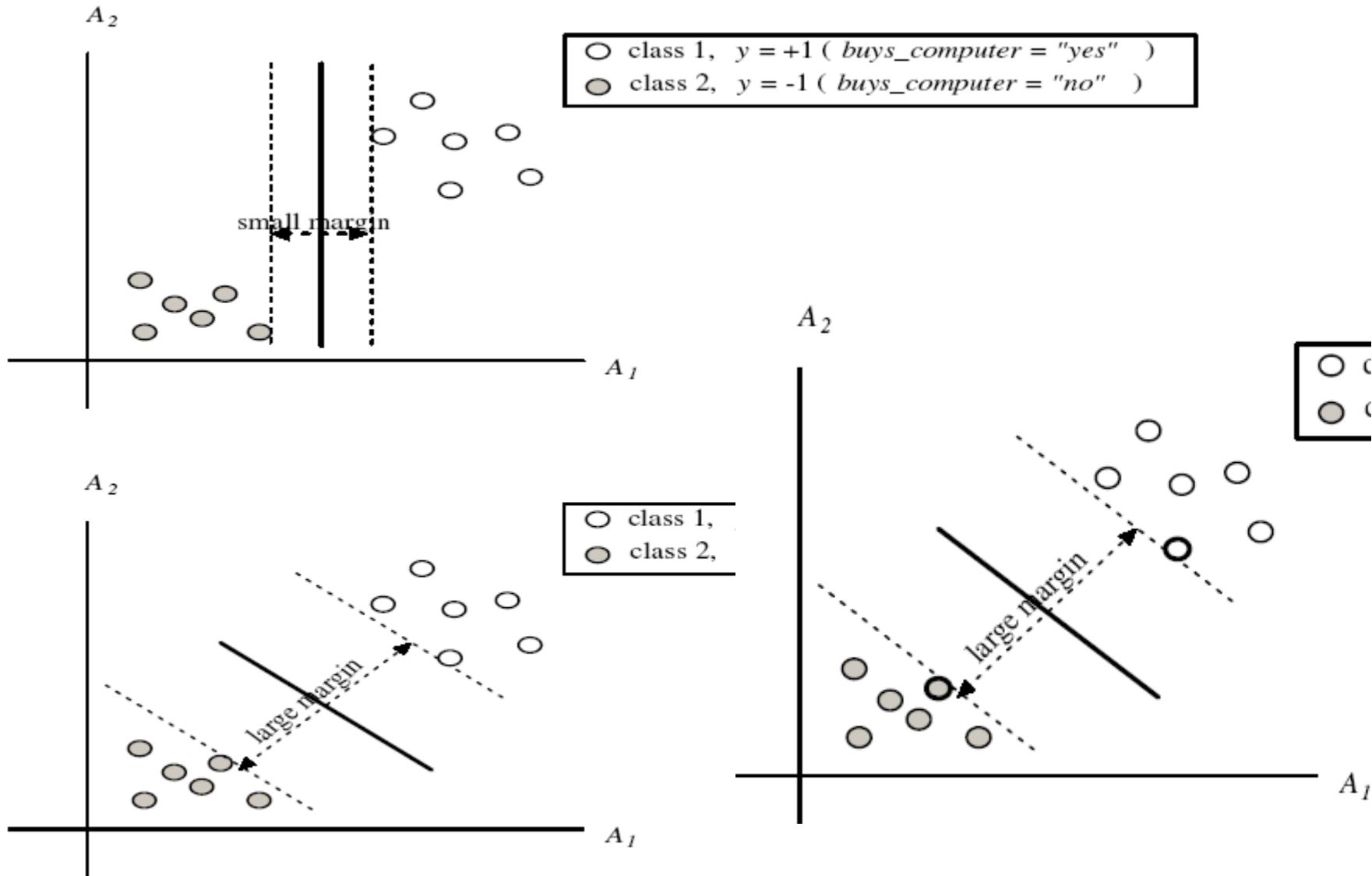
SVM—Support Vector Machines

- A relatively new classification method for both linear and nonlinear data
- It uses a nonlinear mapping to transform the original training data into a higher dimension.
- With the new dimension, it searches for the linear optimal separating **hyperplane** (i.e., “decision boundary”).
- With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane.
- SVM finds this hyperplane using **support vectors** (“essential” training tuples) and **margins** (defined by the support vectors).
- Features: training can be slow but accuracy is high owing to their ability to model complex nonlinear decision boundaries (margin maximization)
- Used for: classification and numeric prediction
- Applications: handwritten digit recognition, object recognition, speaker identification, benchmarking time-series prediction tests

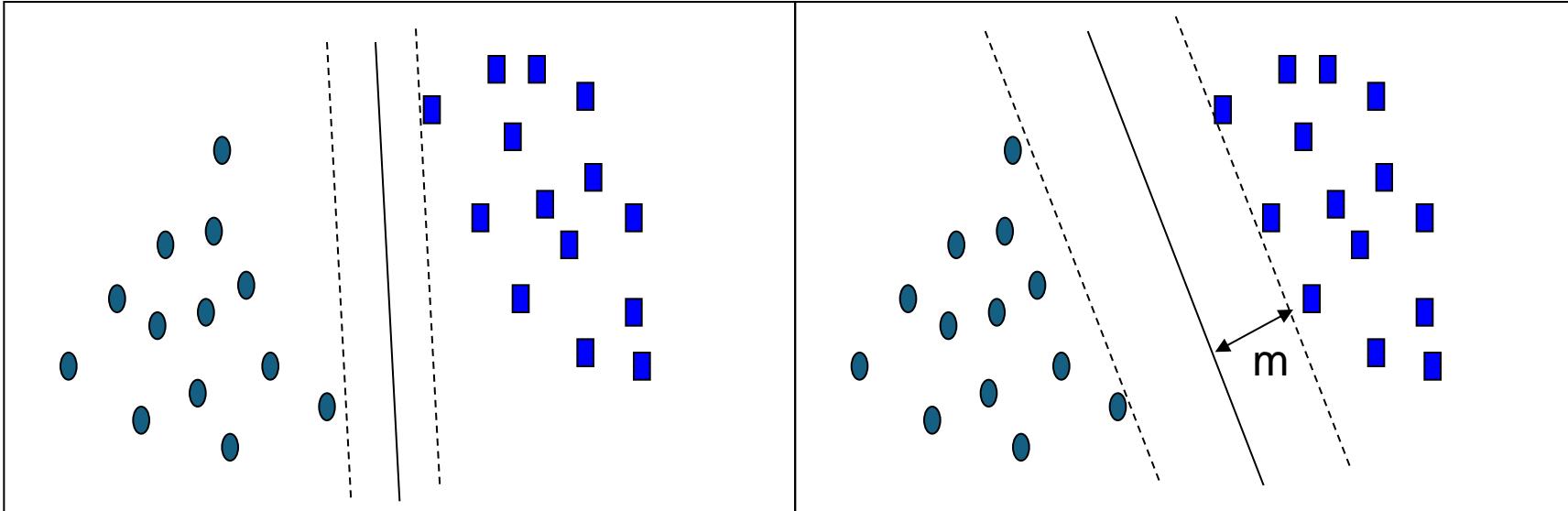
SVM—General Philosophy



SVM—Margins and Support Vectors



SVM—When Data Is Linearly Separable



Let data D be $(\mathbf{X}_1, y_1), \dots, (\mathbf{X}_{|D|}, y_{|D|})$, where \mathbf{X}_i is the set of training tuples associated with the class labels y_i

There are infinite lines (hyperplanes) separating the two classes but we want to find the best one (the one that minimizes classification error on unseen data)

*SVM searches for the hyperplane with the largest margin, i.e., **maximum marginal hyperplane (MMH)***

SVM—Linearly Separable

- A separating hyperplane can be written as

$$\mathbf{w} \bullet \mathbf{x} + b = 0$$

where $\mathbf{w}=\{w_1, w_2, \dots, w_n\}$ is a weight vector and b a scalar (bias)

- For 2-D it can be written as

$$w_0 + w_1 x_1 + w_2 x_2 = 0$$

- The hyperplane defining the sides of the margin:

$$H_1: w_0 + w_1 x_1 + w_2 x_2 \geq 1 \quad \text{for } y_i = +1, \text{ and}$$

$$H_2: w_0 + w_1 x_1 + w_2 x_2 \leq -1 \quad \text{for } y_i = -1$$

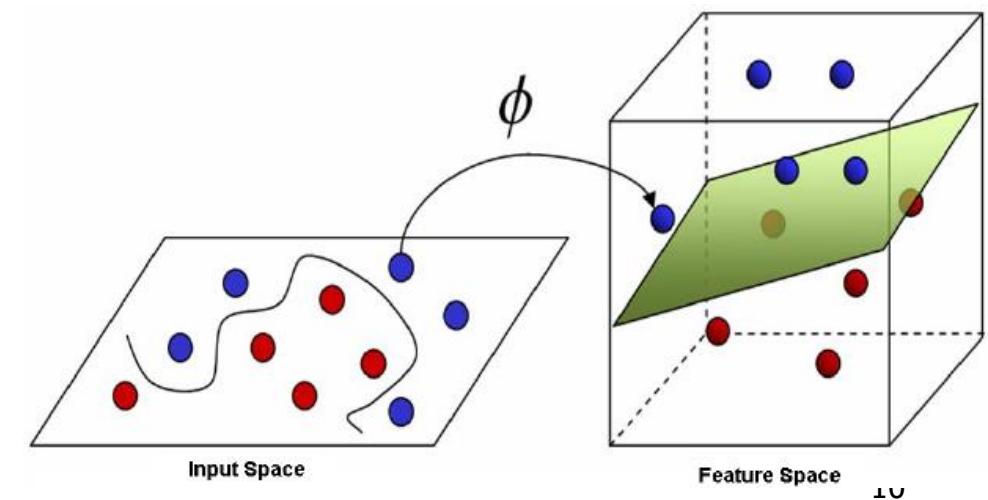
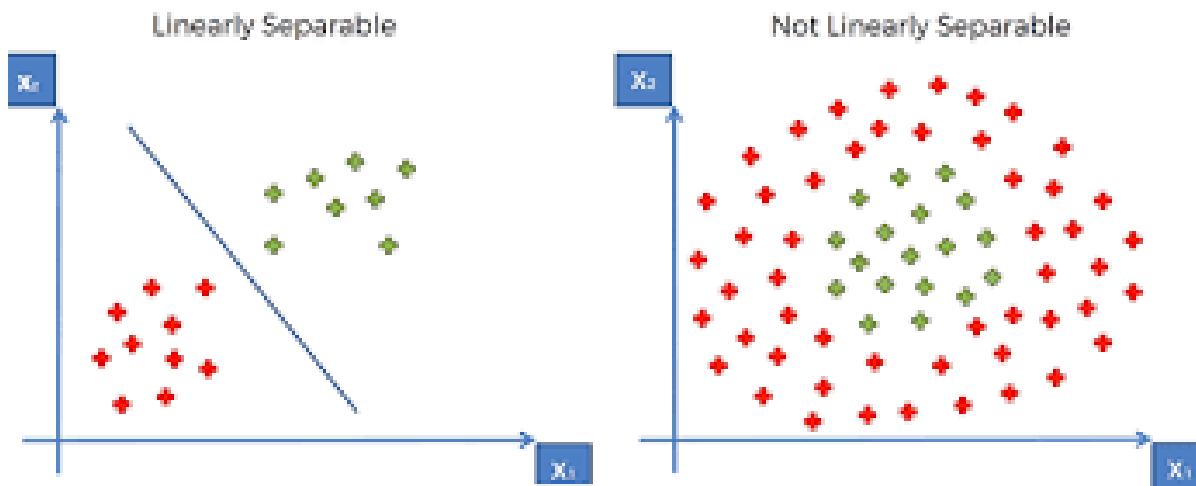
- Any training tuples that fall on hyperplanes H_1 or H_2 (i.e., the sides defining the margin) are **support vectors**
- This becomes a **constrained (convex) quadratic optimization** problem: Quadratic objective function and linear constraints \rightarrow *Quadratic Programming (QP)* \rightarrow Lagrangian multipliers

Why Is SVM Effective on High Dimensional Data?

- The **complexity** of trained classifier is characterized by the # of support vectors rather than the dimensionality of the data
- The **support vectors** are the essential or critical training examples —they lie closest to the decision boundary (MMH)
- If all other training examples are removed and the training is repeated, the same separating hyperplane would be found
- The number of support vectors found can be used to compute an (upper) bound on the expected error rate of the SVM classifier, which is independent of the data dimensionality
- Thus, an SVM with a small number of support vectors can have good generalization, even when the dimensionality of the data is high

SVM—Linearly Inseparable

- Transform the original input data into a higher dimensional space.
- Non-linear SVMs use kernel functions to transform data into higher-dimensional spaces, allowing for the creation of complex, non-linear decision boundaries that are not possible with linear SVMs, enabling effective classification of non-linearly separable data.
- Instead of computing the dot product on the transformed data, it is math. equivalent to applying a kernel function $K(\mathbf{X}_i, \mathbf{X}_j)$ to the original data, i.e., $K(\mathbf{X}_i, \mathbf{X}_j) = \Phi(\mathbf{X}_i) \Phi(\mathbf{X}_j)$
- Search for a linear separating hyperplane in the new space.



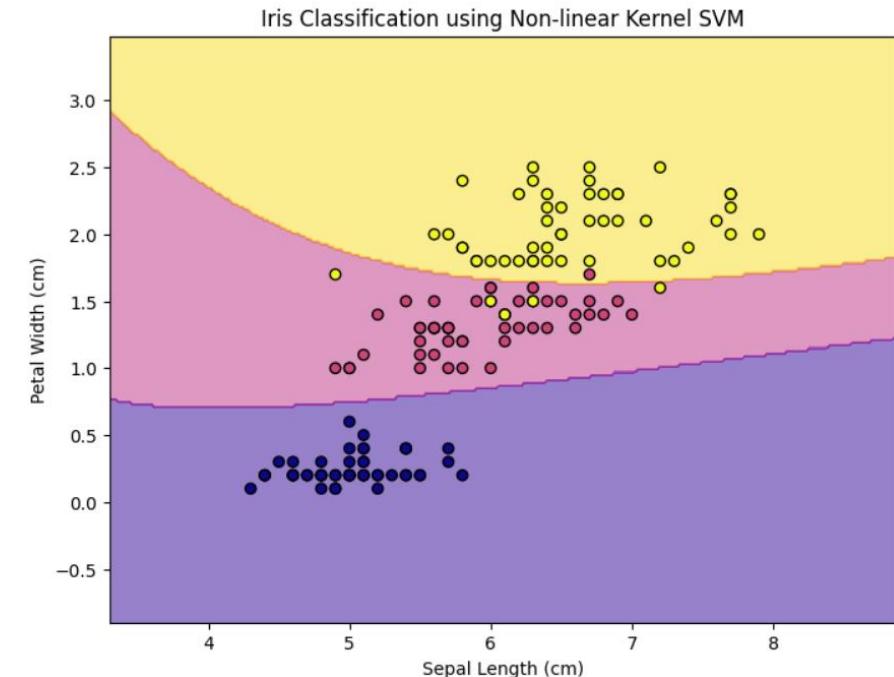
SVM: Different Kernel functions

Typical Kernel Functions

Polynomial kernel of degree h : $K(X_i, X_j) = (X_i \cdot X_j + 1)^h$

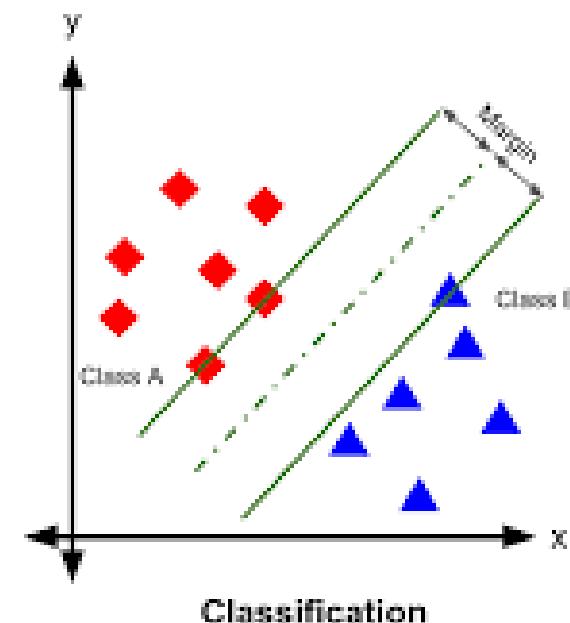
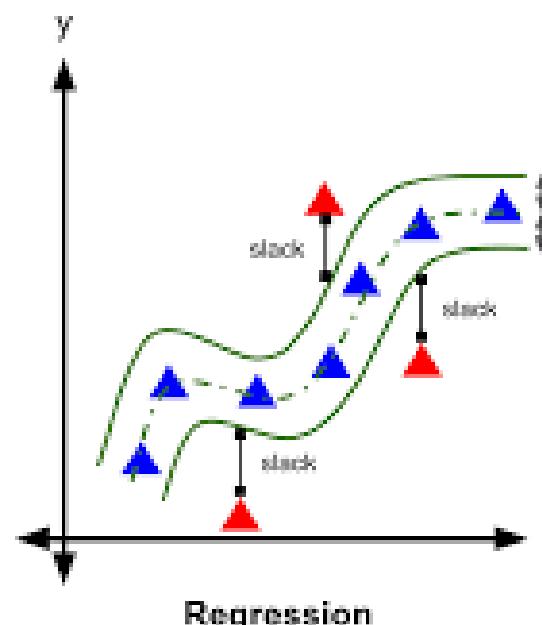
Gaussian radial basis function kernel : $K(X_i, X_j) = e^{-\|X_i - X_j\|^2 / 2\sigma^2}$

Sigmoid kernel : $K(X_i, X_j) = \tanh(\kappa X_i \cdot X_j - \delta)$



SVM can also be used for

- classifying multiple (> 2) classes and
- regression analysis



Model Evaluation and Selection

- Evaluation metrics: How can we measure accuracy? Other metrics to consider?
- Use **validation/test set** of class-labeled tuples instead of training set when assessing accuracy
- Methods for estimating a classifier's accuracy:
 - Holdout method, random subsampling
 - Cross-validation
 - Bootstrap

Classifier Evaluation Metrics: Confusion Matrix

Confusion Matrix:

Actual class\Predicted class	C_1	$\neg C_1$
C_1	True Positives (TP)	False Negatives (FN)
$\neg C_1$	False Positives (FP)	True Negatives (TN)

Example of Confusion Matrix:

Actual class\Predicted class	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	6954	46	7000
buy_computer = no	412	2588	3000
Total	7366	2634	10000

- Given m classes, an entry, $CM_{i,j}$ in a **confusion matrix** indicates # of tuples in class i that were labeled by the classifier as class j
- May have extra rows/columns to provide totals

Classifier Evaluation Metrics

- **Classifier Accuracy**, or recognition rate: percentage of test set tuples that are correctly classified.

$$\text{Accuracy} = (\text{TP} + \text{TN})/\text{All}$$

- **Error rate** = $1 - \text{accuracy} = (\text{FP} + \text{FN})/\text{All}$

- **Precision**: exactness – what % of tuples that the classifier labeled as positive are actually positive.
- **Recall**: completeness – what % of positive tuples did the classifier label as positive?
- Perfect score is 1.0

$$precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- Inverse relationship between precision & recall
- **F measure (F₁- score)**: harmonic mean of precision and recall.

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Classifier Evaluation Metrics

Classification Metrics Examples

Actual	Predicted		Row Totals
	Positive	Negative	
Positive	60	10	70
Negative	5	25	30
Col Totals	65	35	100

$$\text{Precision} = \frac{60}{65} = 0.923$$

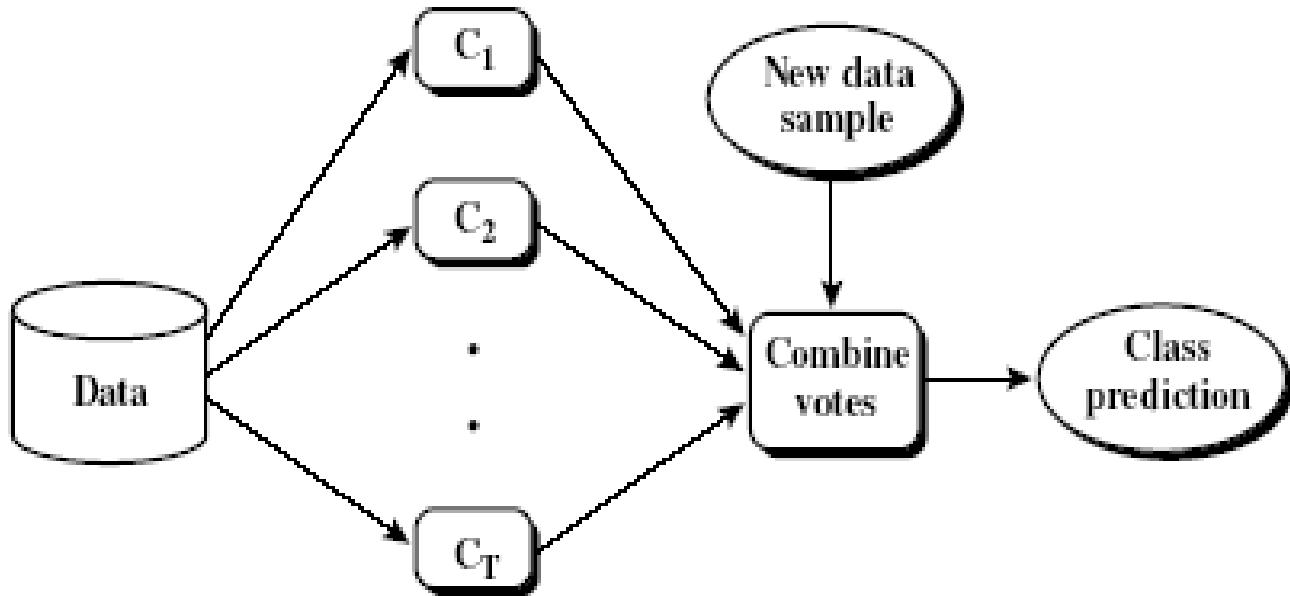
$$\text{Accuracy} = \frac{85}{100} = 85\%$$

$$\begin{aligned}\text{Recall} &= \frac{60}{70} = 0.857 \\ \text{Specificity} &= \frac{25}{30} = 0.833 \\ \text{Error} &= \frac{15}{100} = 15\% \\ F &= 2 * \frac{0.857 * 0.923}{0.857 + 0.923} = 0.889\end{aligned}$$

Holdout & Cross-Validation Methods

- **Holdout method**
 - Given data is randomly partitioned into two independent sets
 - Training set (e.g., 2/3) for model construction
 - Test set (e.g., 1/3) for accuracy estimation
 - Random sampling: a variation of holdout
 - Repeat holdout k times, accuracy = avg. of the accuracies obtained
- **Cross-validation** (k -fold, where $k = 10$ is most popular)
 - Randomly partition the data into k *mutually exclusive* subsets, each approximately equal size
 - At i -th iteration, use D_i as test set and others as training set
 - Leave-one-out: k folds where $k = \#$ of tuples, for small sized data
 - ***Stratified cross-validation***: folds are stratified so that class dist. in each fold is approx. the same as that in the initial data

Ensemble Methods: Increasing the Accuracy



- Ensemble methods
 - Use a combination of models to increase accuracy
 - Combine a series of k learned models, M_1, M_2, \dots, M_k , with the aim of creating an improved model M^*
- Popular ensemble methods
 - Bagging: averaging the prediction over a collection of classifiers
 - Boosting: weighted vote with a collection of classifiers
 - Ensemble: combining a set of heterogeneous classifiers

Random Forest (Breiman 2001)

- Random Forest:
 - Each classifier in the ensemble is a *decision tree* classifier and is generated using a random selection of attributes at each node to determine the split
 - During classification, each tree votes and the most popular class is returned
- Two Methods to construct Random Forest:
 - Forest-RI (*random input selection*): Randomly select, at each node, F attributes as candidates for the split at the node. The CART methodology is used to grow the trees to maximum size
 - Forest-RC (*random linear combinations*): Creates new attributes (or features) that are a linear combination of the existing attributes (reduces the correlation between individual classifiers)
- Comparable in accuracy to Adaboost, but more robust to errors and outliers
- Insensitive to the number of attributes selected for consideration at each split, and faster than bagging or boosting

Supervised Learning

k-Nearest Neighbor Classifier

- k -NN classification rule is to assign to a test sample the majority category label of its k nearest training samples.
- In practice, k is usually chosen to be odd, so as to avoid ties
- The $k = 1$ rule is generally called the nearest-neighbor classification rule.

k-Nearest Neighbor Classifier

- The model is **not trained** beforehand, it runs at a time of execution to find the query output. There's no point in training the model earlier as an input of the model is training data, **hyperparameter (k)**, and a query point (x_q).
- **Steps of K-NN classifier:**
 - **Distance Calculation:** KNN calculates the distance between a new data point and all other data points in the dataset.
 - **Finding Nearest Neighbors:** It then identifies the 'k' nearest neighbors to the new data point.
 - **Prediction:** For classification, the new data point is assigned to the class that is most common among its 'k' nearest neighbors. For regression, the predicted value is the average of the values of its 'k' nearest neighbors.

Nearest-Neighbor Classifiers: Issues

- The value of k , the number of nearest neighbors to retrieve
- Choice of Distance Metric to compute distance between records
- Computational complexity
 - Size of training set
 - Dimension of data

Euclidean

$$\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

Manhattan

$$\sum_{i=1}^k |x_i - y_i|$$

Minkowski

$$\left(\sum_{i=1}^k (|x_i - y_i|)^q \right)^{1/q}$$

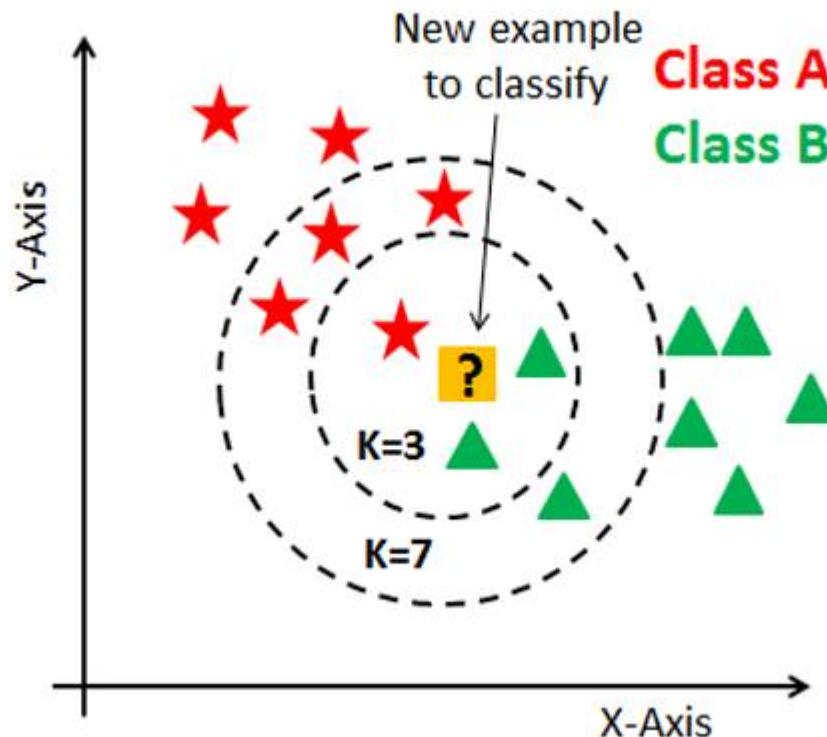
Value of K

- Choosing the value of k:
 - If k is too small, sensitive to noise points
 - If k is too large, neighborhood may include points from other classes

Rule of thumb:

$$K = \sqrt{N}$$

N: number of training points



k-Nearest Neighbor Applications

KNN is used in various **applications**, including:

- Image Recognition: Identifying objects in images.
- Recommendation Systems: Suggesting products or content based on user preferences.
- Fraud Detection: Identifying fraudulent transactions.

Advantages:

- Simplicity: KNN is a straightforward algorithm to understand and implement.
- Versatility: It can be used for both classification and regression tasks.
- No Training Phase: It doesn't require a separate training phase, making it efficient for certain datasets.

Nearest Neighbour : Computational Complexity

- Expensive
 - To determine the nearest neighbour of a query point q , must compute the distance to all N training examples
 - + Pre-sort training examples into fast data structures (kd-trees)
 - + Compute only an approximate distance (LSH)
 - + Remove redundant data (condensing)
- Storage Requirements
 - Must store all training data \mathbf{P}
 - + Remove redundant data (condensing)
 - Pre-sorting often increases the storage requirements
- High Dimensional Data
 - “Curse of Dimensionality”
 - Required amount of training data increases exponentially with dimension
 - Computational cost also increases dramatically
 - Partitioning techniques degrade to linear search in high dimension

Supervised Learning

Clustering

- Cluster Analysis: Basic Concepts
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Evaluation of Clustering
- Summary



What is Cluster Analysis?

- Cluster: A collection of data objects
 - similar (or related) to one another within the same group
 - dissimilar (or unrelated) to the objects in other groups
- Cluster analysis (or *clustering*, *data segmentation*, ...)
 - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
- **Unsupervised learning**: no predefined classes (i.e., *learning by observations* vs. learning by examples: supervised)
- Typical applications
 - As a **stand-alone tool** to get insight into data distribution
 - As a **preprocessing step** for other algorithms

Clustering for Data Understanding and Applications

- **Biology:** taxonomy of living things: kingdom, phylum, class, order, family, genus and species
- **Information retrieval:** document clustering
- **Land use:** Identification of areas of similar land use in an earth observation database
- **Marketing:** Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- **City-planning:** Identifying groups of houses according to their house type, value, and geographical location
- **Earth-quake studies:** Observed earthquake epicenters should be clustered along continent faults
- **Climate:** understanding earth climate, find patterns of atmospheric and ocean
- **Economic Science:** market research

Clustering as a Preprocessing Tool (Utility)

- Summarization:
 - Preprocessing for regression, PCA, classification, and association analysis
- Compression:
 - Image processing: vector quantization
- Finding K-nearest Neighbors
 - Localizing search to one or a small number of clusters
- Outlier detection
 - Outliers are often viewed as those “far away” from any cluster

Quality: What Is Good Clustering?

- A good clustering method will produce high quality clusters
 - high intra-class similarity: cohesive within clusters
 - low inter-class similarity: distinctive between clusters
- The quality of a clustering method depends on
 - the similarity measure used by the method
 - its implementation, and
 - Its ability to discover some or all of the hidden patterns

Measure the Quality of Clustering

- Dissimilarity/Similarity metric
 - Similarity is expressed in terms of a distance function, typically metric: $d(i, j)$
 - The definitions of **distance functions** are usually rather different for interval-scaled, boolean, categorical, ordinal ratio, and vector variables
 - Weights should be associated with different variables based on applications and data semantics
- Quality of clustering:
 - There is usually a separate “quality” function that measures the “goodness” of a cluster.
 - It is hard to define “similar enough” or “good enough”
 - The answer is typically highly subjective

Requirements and Challenges

- Scalability
 - Clustering all the data instead of only on samples
- Ability to deal with different types of attributes
 - Numerical, binary, categorical, ordinal, linked, and mixture of these
- Constraint-based clustering
 - User may give inputs on constraints
 - Use domain knowledge to determine input parameters
- Interpretability and usability
- Others
 - Discovery of clusters with arbitrary shape
 - Ability to deal with noisy data
 - Incremental clustering and insensitivity to input order
 - High dimensionality

Major Clustering Approaches (I)

- Partitioning approach:
 - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
 - Typical methods: k-means, k-medoids, CLARANS
- Hierarchical approach:
 - Create a hierarchical decomposition of the set of data (or objects) using some criterion
 - Typical methods: Diana, Agnes, BIRCH, CAMELEON
- Density-based approach:
 - Based on connectivity and density functions
 - Typical methods: DBSACN, OPTICS, DenClue
- Grid-based approach:
 - based on a multiple-level granularity structure
 - Typical methods: STING, WaveCluster, CLIQUE

Partitioning Algorithms: Basic Concept

- Partitioning method: Partitioning a database D of n objects into a set of k clusters, such that the sum of squared distances is minimized (where c_i is the centroid or medoid of cluster C_i)

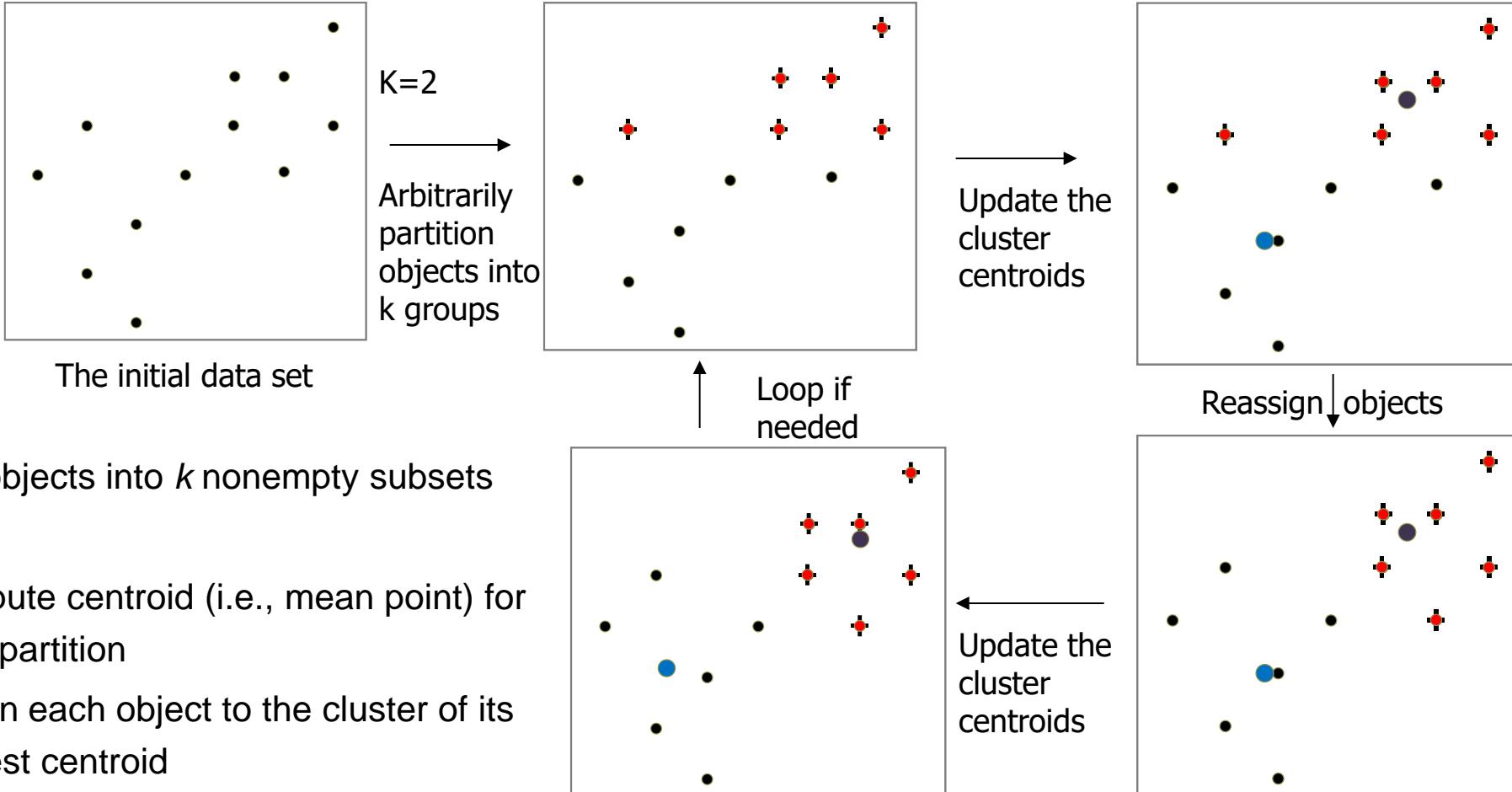
$$E = \sum_{i=1}^k \sum_{p \in C_i} (p - c_i)^2$$

- Given k , find a partition of k clusters that optimizes the chosen partitioning criterion
 - Global optimal: exhaustively enumerate all partitions
 - Heuristic methods: *k-means* and *k-medoids* algorithms
 - *k-means* (MacQueen'67, Lloyd'57/'82): Each cluster is represented by the center of the cluster
 - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster

The *K*-Means Clustering Method

- Given k , the *k-means* algorithm is implemented in four steps:
 - Partition objects into k nonempty subsets
 - Compute seed points as the centroids of the clusters of the current partitioning (the centroid is the center, i.e., *mean point*, of the cluster)
 - Assign each object to the cluster with the nearest seed point
 - Go back to Step 2, stop when the assignment does not change

An Example of K -Means Clustering



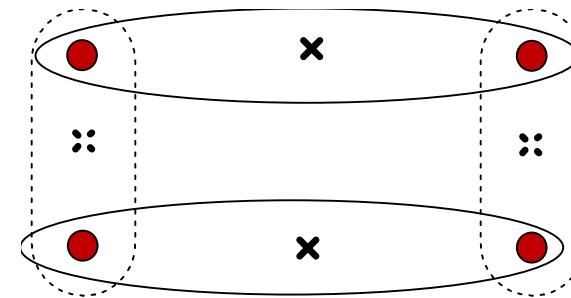
- Partition objects into k nonempty subsets
- Repeat
 - Compute centroid (i.e., mean point) for each partition
 - Assign each object to the cluster of its nearest centroid
- Until no change

Comments on the *K-Means* Method

- Strength: *Efficient*: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations.
 - Normally, $k, t \ll n$.
 - Comparing: PAM: $O(k(n-k)^2)$, CLARA: $O(ks^2 + k(n-k))$
- Comment: Often terminates at a *local optimal*.
- Weakness
 - Applicable only to objects in a continuous n-dimensional space
 - Using the k-modes method for categorical data
 - In comparison, k-medoids can be applied to a wide range of data
 - Need to specify k , the *number* of clusters, in advance (there are ways to automatically determine the best k (see Hastie et al., 2009))
 - Sensitive to noisy data and *outliers*
 - Not suitable to discover clusters with *non-convex shapes*

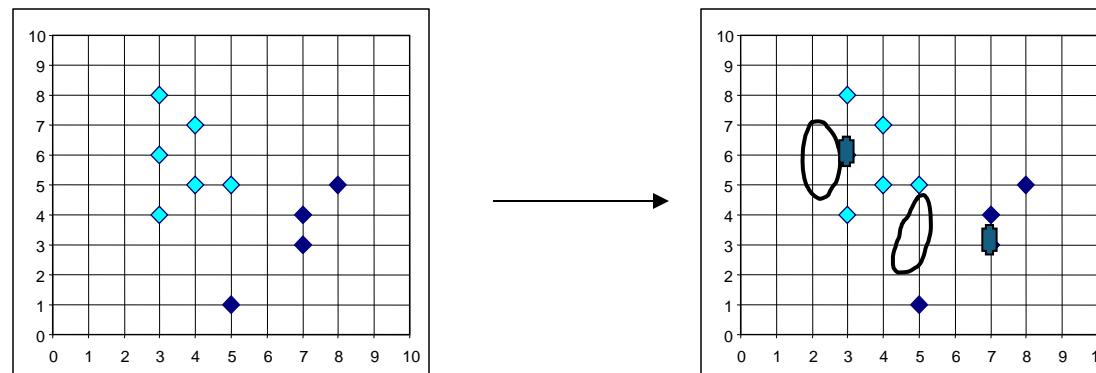
Variations of the *K-Means* Method

- Most of the variants of the *k-means* which differ in
 - Selection of the initial *k* means
 - Dissimilarity calculations
 - Strategies to calculate cluster means
- Handling categorical data: *k-modes*
 - Replacing means of clusters with modes
 - Using new dissimilarity measures to deal with categorical objects
 - Using a frequency-based method to update modes of clusters
 - A mixture of categorical and numerical data: *k-prototype* method



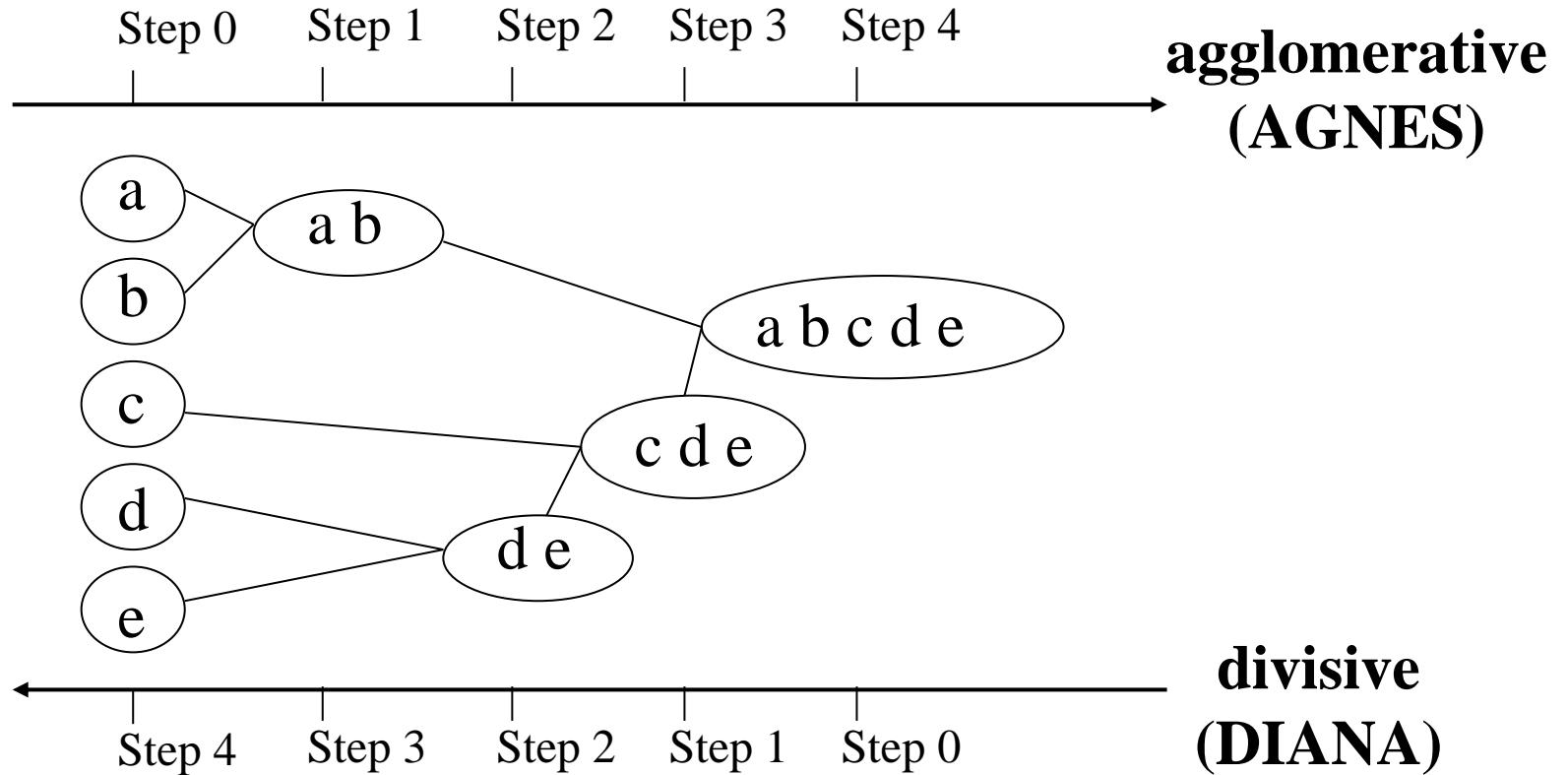
What Is the Problem of the K-Means Method?

- The k-means algorithm is sensitive to outliers !
 - Since an object with an extremely large value may substantially distort the distribution of the data
- K-Medoids: Instead of taking the **mean** value of the object in a cluster as a reference point, **medoids** can be used, which is the **most centrally located** object in a cluster

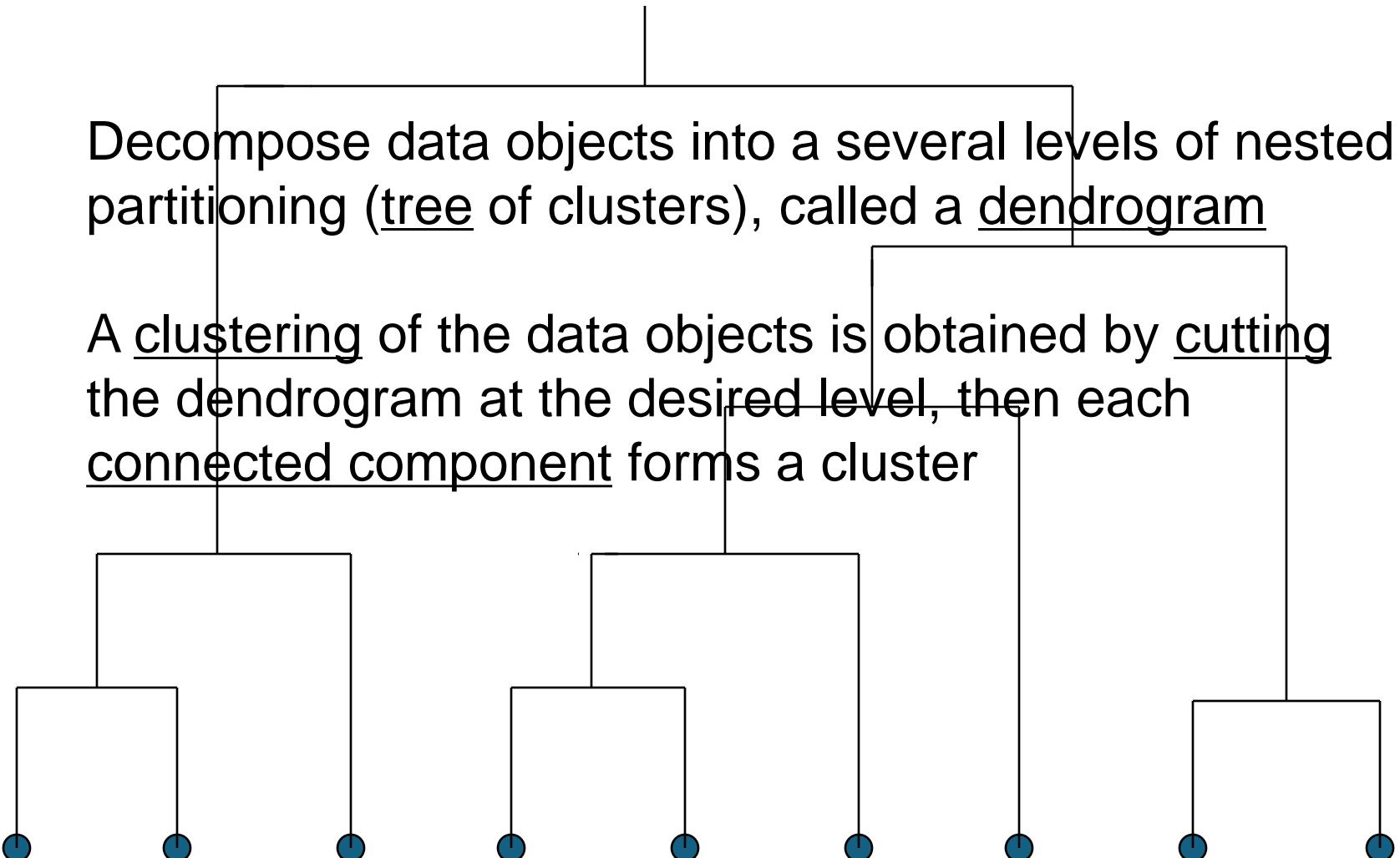


Hierarchical Clustering

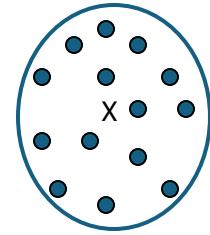
- Use distance matrix as clustering criteria. This method does not require the number of clusters k as an input, but needs a termination condition



Dendrogram: Shows How Clusters are Merged



Distance between Clusters

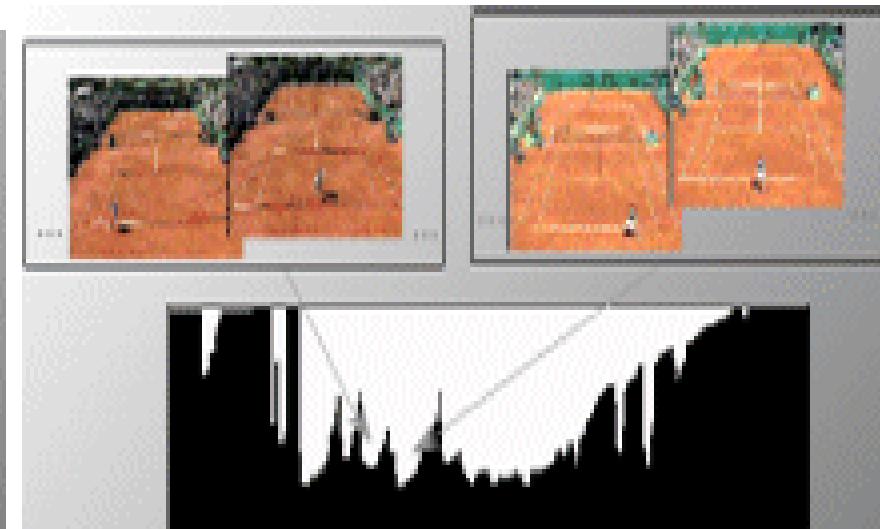
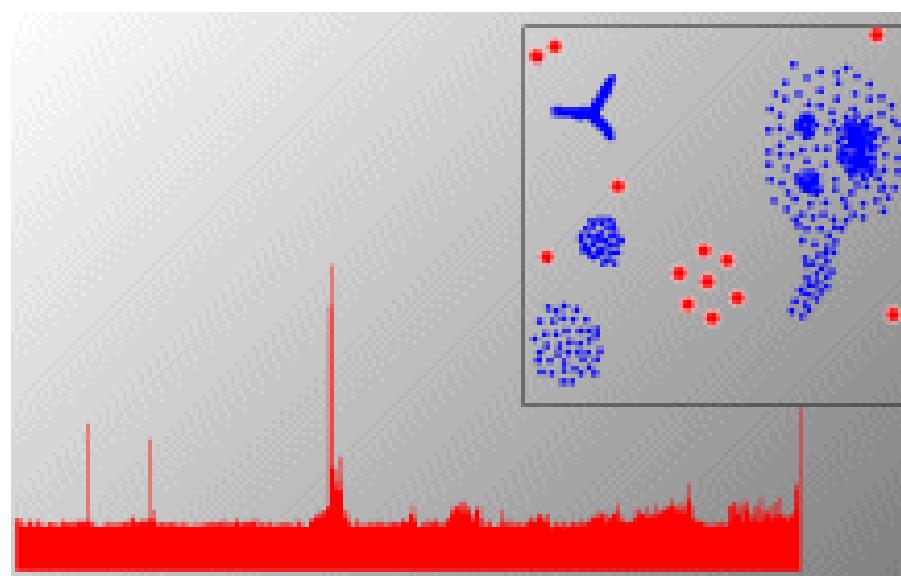
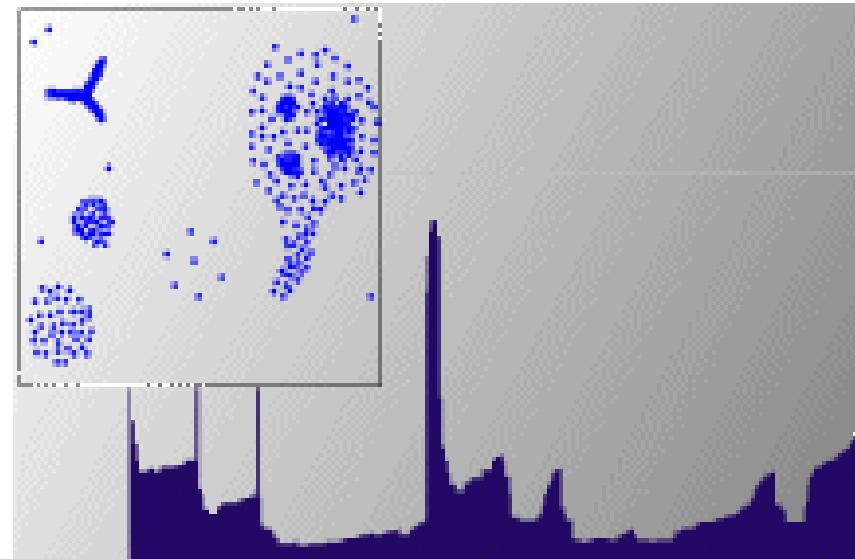
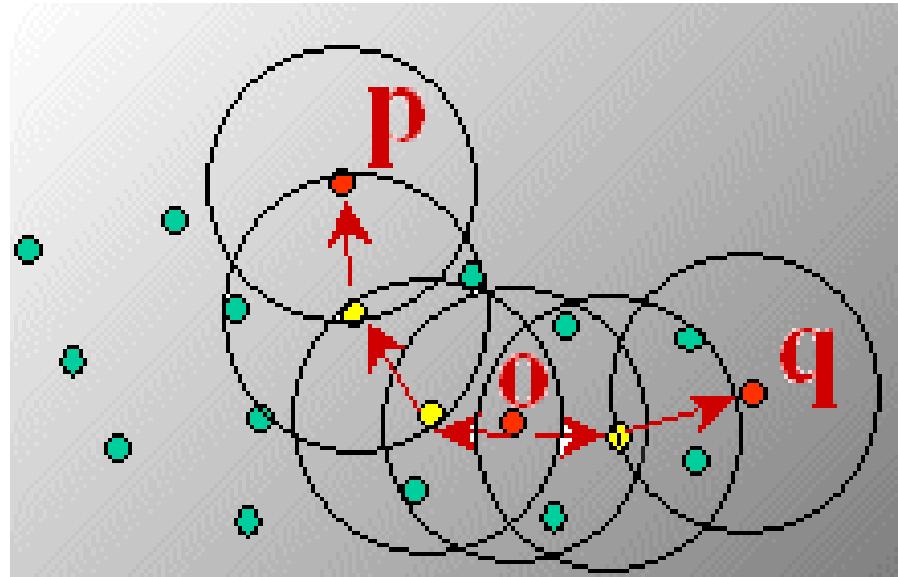


- Single link: smallest distance between an element in one cluster and an element in the other, i.e., $\text{dist}(K_i, K_j) = \min(t_{ip}, t_{jq})$
- Complete link: largest distance between an element in one cluster and an element in the other, i.e., $\text{dist}(K_i, K_j) = \max(t_{ip}, t_{jq})$
- Average: avg distance between an element in one cluster and an element in the other, i.e., $\text{dist}(K_i, K_j) = \text{avg}(t_{ip}, t_{jq})$
- Centroid: distance between the centroids of two clusters, i.e., $\text{dist}(K_i, K_j) = \text{dist}(C_i, C_j)$
- Medoid: distance between the medoids of two clusters, i.e., $\text{dist}(K_i, K_j) = \text{dist}(M_i, M_j)$
 - Medoid: a chosen, centrally located object in the cluster

Density-Based Clustering Methods

- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
 - Discover clusters of arbitrary shape
 - Handle noise
 - One scan
 - Need density parameters as termination condition
- Several interesting studies:
 - DBSCAN: Ester, et al. (KDD'96)
 - OPTICS: Ankerst, et al (SIGMOD'99).
 - DENCLUE: Hinneburg & D. Keim (KDD'98)
 - CLIQUE: Agrawal, et al. (SIGMOD'98) (more grid-based)

Density-Based Clustering: OPTICS & Its Applications



Measuring Clustering Quality

- Two methods: extrinsic vs. intrinsic
- Extrinsic: supervised, i.e., the ground truth is available
 - Compare a clustering against the ground truth using certain clustering quality measure
 - Ex. BCubed precision and recall metrics
- Intrinsic: unsupervised, i.e., the ground truth is unavailable
 - Evaluate the goodness of a clustering by considering how well the clusters are separated, and how compact the clusters are
 - Ex. Silhouette coefficient

Measuring Clustering Quality: Extrinsic Methods

- Clustering quality measure: $Q(C, C_g)$, for a clustering C given the ground truth C_g .
- Q is good if it satisfies the following **4** essential criteria
 - Cluster homogeneity: the purer, the better
 - Cluster completeness: should assign objects belong to the same category in the ground truth to the same cluster
 - Rag bag: putting a heterogeneous object into a pure cluster should be penalized more than putting it into a *rag bag* (i.e., “miscellaneous” or “other” category)
 - Small cluster preservation: splitting a small category into pieces is more harmful than splitting a large category into pieces