

Math Basics

Session 1: (Statistics Fundamentals)

Agenda:

1	Introduction to Statistics (30min)
2	Descriptive Statistics (1 hour)
3	Probability Basics (1 hour)
4	Introduction to Inferential Statistics (30 mins)

Introduction to Statistics (30min)

Importance of Statistics in AI and Machine Learning

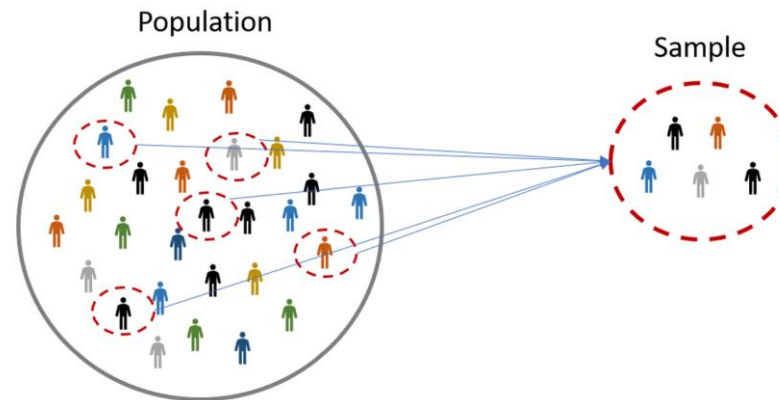
Why Learn Statistics for AI?

- **Data Analysis:** Statistics provide the tools to analyze and interpret data, which is the backbone of AI. Understanding data distributions, variability, and patterns is crucial for building robust models.
- **Model Evaluation:** In AI, especially in machine learning, evaluating the performance of models (e.g., accuracy, precision, recall) relies on statistical measures. Without a strong understanding of statistics, interpreting these metrics can be challenging.
- **Decision Making:** AI models often make decisions based on probabilities. For example, a classification model might predict the likelihood of an email being spam. Understanding probability and statistical inference is key to trusting and explaining these decisions.
- **Real-World Applications:** From predicting customer behavior in marketing to diagnosing diseases in healthcare, statistics provide the foundation for analyzing data and making informed decisions.

Basic Statistical Definitions

Population vs. Sample

- **Population:** The entire group of individuals or items that you're interested in studying. For example, if you want to study the average height of adult males in a country, the population would be **all** adult males in that country.
- **Sample:** A subset of the population that is used to represent the whole. For instance, measuring the height of **1,000 adult males** from different regions could be a sample representing the entire population. Sampling is often done because it is impractical or impossible to study the whole population.

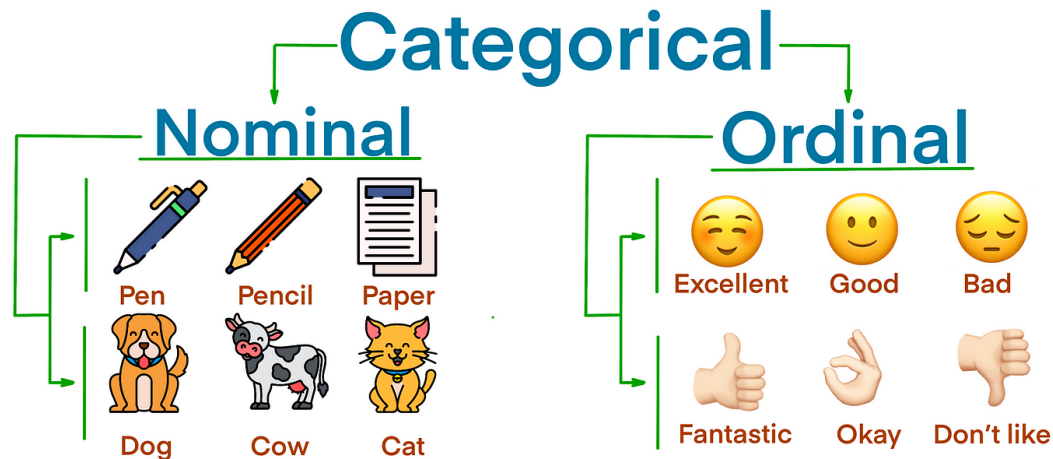


Types of Variables

Categorical (Qualitative) Variables

These variables represent categories or groups and can be either:

- **Nominal: No inherent order among categories.** Examples include colors (red, blue, green), gender (male, female).
- **Ordinal: There is an inherent order, but the difference between levels is not quantifiable.** Examples include customer satisfaction ratings (poor, fair, good, excellent), education levels (high school, bachelor's, master's).



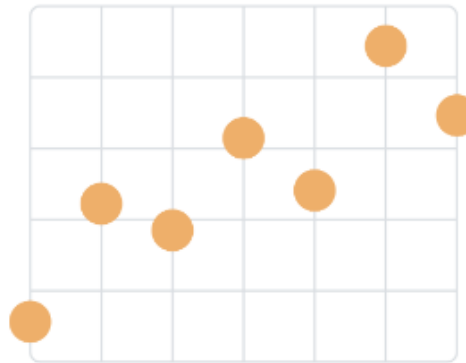
Types of Variables

Numerical (Quantitative) Variables

These variables represent measurable quantities and can be either:

- **Discrete:** Countable, finite values. Examples include the number of students in a class, the number of cars in a parking lot.
- **Continuous:** Infinite possibilities within a range. Examples include height, weight, temperature.

DISCRETE DATA



CONTINUOUS DATA



Real-World Relevance and Applications

Examples of Statistics in AI Applications

- **Fraud Detection:** Analyzing transaction data to identify patterns that indicate fraudulent activity. Here, statistics help to detect anomalies that deviate from normal behavior.
- **Recommendation Systems:** E-commerce platforms like Amazon use statistics to analyze customer preferences and recommend products based on what similar customers have purchased.
- **Healthcare Analytics:** Predicting the likelihood of diseases based on patient data. Statistical models can assess risk factors and provide personalized healthcare recommendations.



Descriptive Statistics (1 hour)

Descriptive Statistics

Descriptive statistics help provide a clear understanding of data through numerical calculations, graphs, and tables. This is a crucial step before conducting any further statistical analysis or building machine learning models.

Cases vs. Variables

Cases

- A case represents an individual entity or subject in a dataset on which measurements or observations are made. It can be a person, a country, an event, or any other subject of study.

Variables

- A variable is a characteristic or attribute that can be measured or observed for each case. Variables are features or properties that describe the cases.

		Variables			
		<div></div>			
		Store	Sales	Customers	Refunds
Cases	1	14	9	3	
	2	19	13	4	
	3	22	19	4	
	4	24	20	3	
	5	29	26	8	
	6	40	34	6	

Data Matrix vs. Frequency Table

Data Matrix

A data matrix is a structured table where each row represents a single case (individual data point), and each column represents a variable (characteristic or feature of the cases).

Variables				
sepal length	sepal width	petal length	petal width	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
6.5	3.2	5.1	2	Iris-virginica
6.4	2.7	5.3	1.9	Iris-virginica
6.8	3	5.5	2.1	Iris-virginica
6.7	3.1	4.4	1.4	Iris-versicolor
5.6	3	4.5	1.5	Iris-versicolor
5.8	2.7	4.1	1	Iris-versicolor

Data Matrix vs. Frequency Table

Data Matrix

Purpose

The data matrix is useful for raw data representation where each case's specific details are important. It's a comprehensive way to store all the information but not necessarily the best for summarizing data.

When to Use

Data matrices are used in scenarios where you need to keep all details about each individual case, such as during data collection or when you need to perform case-by-case analysis.

Data Matrix vs. Frequency Table

Frequency Table

A frequency table is a summary of the data that shows how often each value of a variable occurs. It can display frequencies, percentages, and cumulative percentages.

Score	Frequency
50-59	2
60-69	2
70-79	6
80-89	7
90-99	3

Data Matrix vs. Frequency Table

Frequency Table

Purpose

Frequency tables are used to summarize and visualize the distribution of a variable, making it easier to see patterns and understand the data's structure.

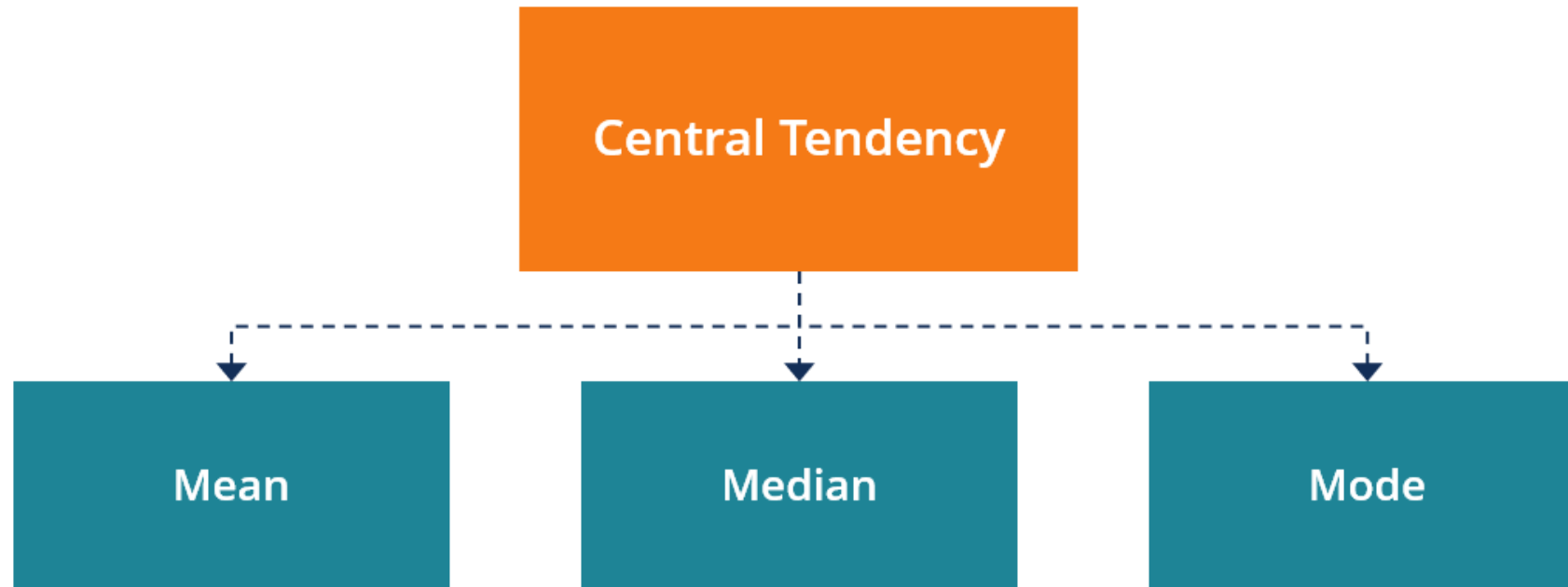
When to Use

When you want to quickly grasp how data is distributed across different categories or ranges, especially when dealing with categorical or quantitative variables.

Descriptive Statistics

Measures of Central Tendency

Understand how to describe the center of a dataset using different measures and learn how to calculate and interpret these values.



Measures of Central Tendency

Mean (Average)

The mean is the sum of all values divided by the number of values. It is a commonly used measure of central tendency.

Formula

For a dataset with values x_1, x_2, \dots, x_n , the mean μ is calculated as:

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

Example

For the dataset $[4, 8, 6, 5, 3, 4]$, the mean is

$$\mu = \frac{4 + 8 + 6 + 5 + 3 + 4}{6} = 5$$

The mean is sensitive to outliers. If the dataset has extreme values (very high or very low), the mean can be misleading

Measures of Central Tendency

Median

The median is the middle value of a dataset when it is ordered from smallest to largest. If the dataset has an even number of values, the median is the average of the two middle values.

Example

For the dataset [4,8,6,5,3,4], first sort the data:[3,4,4,5,6,8]. The median is the average of the two middle numbers, 4 and 5, which is $\frac{4+5}{2} = 4.5$

The median is not affected by **outliers** and provides a **better measure** of central tendency for **skewed distributions**.

Measures of Central Tendency

Mode

The mode is the value that appears most frequently in a dataset.

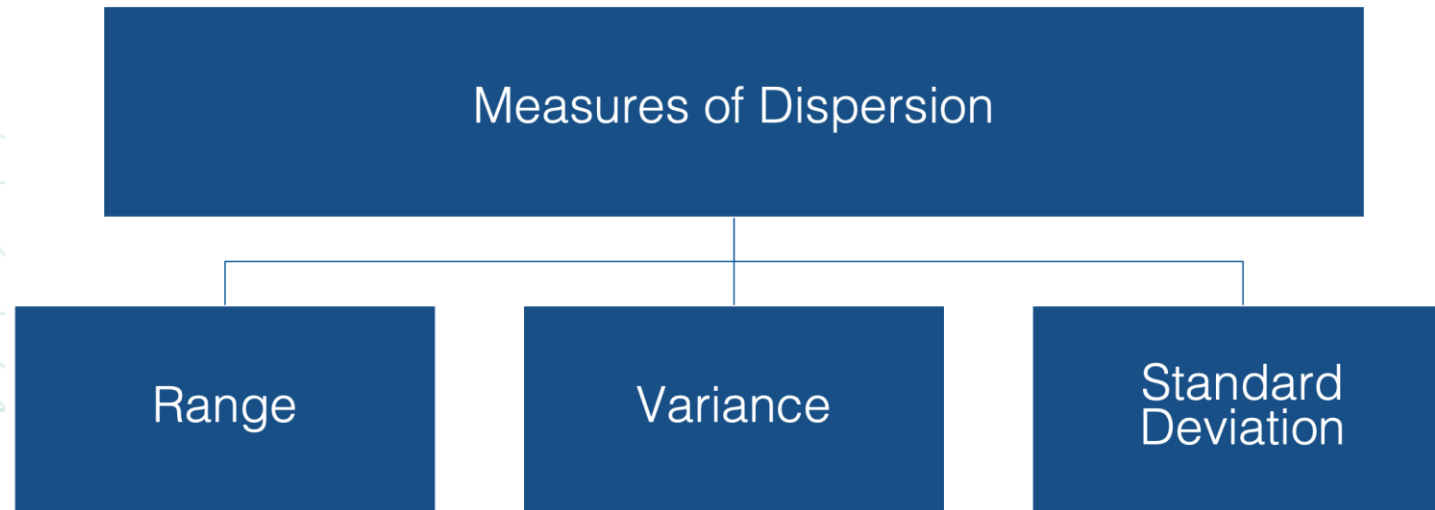
Example

In the dataset $[4, 8, 6, 5, 3, 4]$, the mode is 4 because it appears twice, more than any other number.

The mode is useful for **categorical data** and for identifying the most common item in a dataset. A dataset can have more than one mode (bimodal or multimodal) or no mode at all.

Measures of Dispersion

Learn how to describe the **spread** or **variability** of a dataset using different measures.



Measures of Dispersion

Range

The range is the difference between the **maximum** and **minimum** values in a dataset.

Formula

Range=Max value–Min value.

Example

For the dataset [4,8,6,5,3,4], the range is $8-3=5$.

The range gives a quick sense of the spread of the data but is **highly** sensitive to outliers.

Measures of Dispersion

Interquartile Range (IQR)

The IQR is the range between the first quartile (Q1) and the third quartile (Q3). It measures the spread of the middle 50% of the data, effectively capturing the central tendency without being influenced by extreme values or outliers. The IQR is calculated as

$$IQR = Q3 - Q1$$

- **Q1 (First Quartile):** The value below which 25% of the data falls.
- **Q3 (Third Quartile):** The value below which 75% of the data falls.

Interquartile Range (IQR)

Why Use IQR?

- **Robustness Against Outliers:** Unlike the range, which considers only the minimum and maximum values, the IQR focuses on the middle 50% of the data. This makes it less sensitive to outliers and extreme values, providing a more reliable measure of dispersion for skewed distributions.
- **Understanding Data Spread:** The IQR helps understand the variability of the central portion of the data. A larger IQR indicates more spread in the middle 50% of the dataset, while a smaller IQR suggests that the data points are closer to the median.

Interquartile Range (IQR)

Example Calculation of IQR

Let's use a small dataset to illustrate: **Dataset: [2,4,4,5,6,8,9]**

1. **Arrange the data in ascending order** (already sorted in this case).
2. **Find Q1 and Q3:**
 - **Q1 (First Quartile):** The median of the first half of the dataset (excluding the median if the number of data points is odd). For [2,4,4], $Q1 = 4$
 - **Q3 (Third Quartile):** The median of the second half of the dataset. For [6,8,9], $Q3 = 8$
3. **Calculate IQR** $IQR = Q3 - Q1 = 8 - 4 = 4$

Measures of Dispersion

Variance

Variance measures the **average squared deviation** of each number from the **mean**. It provides insight into the spread of all data points around the mean.

Formula

For a dataset with values x_1, x_2, \dots, x_n , the variance σ^2 is calculated as
$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$$

Example

For the dataset [4,8,6,5,3,4], calculate the mean ($\mu=5$), then compute the variance:

$$\sigma^2 = \frac{(4-5)^2 + (8-5)^2 + (6-5)^2 + (5-5)^2 + (3-5)^2 + (4-5)^2}{6} = \frac{1+9+1+0+4+1}{6} = 2.67$$

Variance is in squared units, making it less interpretable in the original scale. However, it is fundamental in statistical theory.

Measures of Dispersion

Standard Deviation

The standard deviation is the **square root** of the **variance** and provides a measure of the **average distance** from the mean. It is in the same units as the data, making it more **interpretable**.

Formula: $\sigma = \sqrt{\sigma^2}$

Example

For the dataset [4,8,6,5,3,4], the standard deviation is: $\sigma = \sqrt{2.67} \approx 1.63$

A small standard deviation indicates that the values are close to the mean, while a large standard deviation indicates that the values are spread out over a wider range.

Descriptive Statistics

Data Distribution and Visualization

Learn how to visualize data distribution and understand its shape using graphical representations.



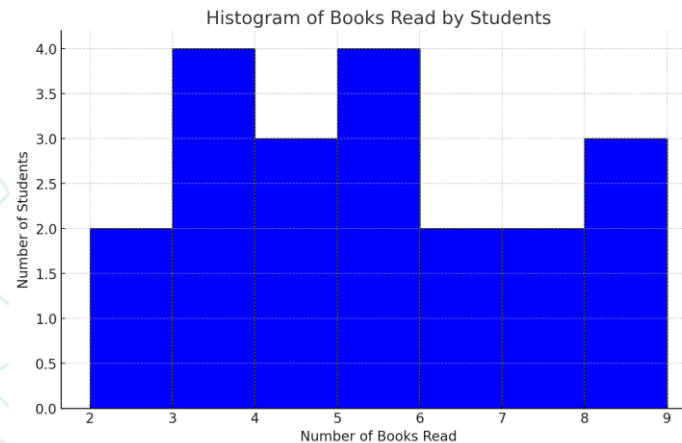
Data Distribution and Visualization

Histograms

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. It shows the frequency distribution of a dataset.

Example

Create a histogram for a dataset representing the number of books read by students in a class.



- Histograms help visualize the shape of the data distribution (e.g., normal, skewed) and identify patterns like bimodality or skewness.

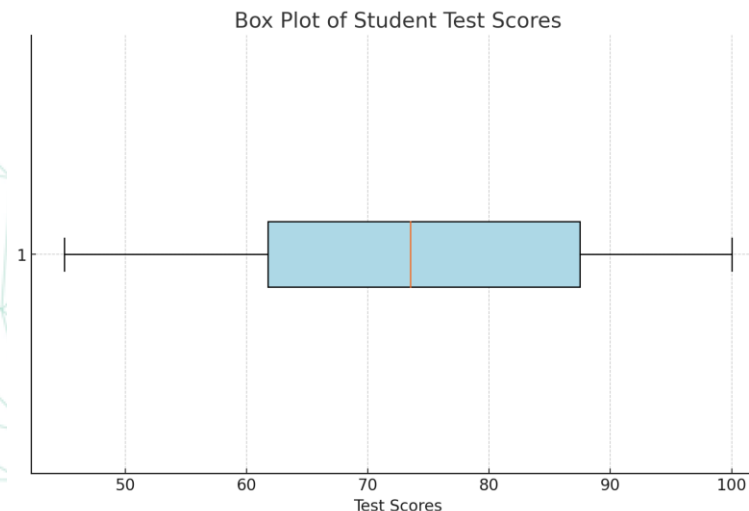
Data Distribution and Visualization

Box Plots (Box-and-Whisker Plots)

A box plot displays the distribution of data based on a five-number summary: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.

Example

Create a box plot for a dataset of test scores to identify the median score, quartiles, and potential outliers.



- Box plots are useful for detecting outliers and understanding the spread and symmetry of the data.

Quartiles

First Quartile (Q1, 25th percentile): The left edge of the box represents the 25th percentile, indicating that 25% of the students scored below this value.

Third Quartile (Q3, 75th percentile): The right edge of the box represents the 75th percentile, showing that 75% of the students scored below this value.

Probability Basics (1 hour)

Probability Basics

fundamental probability concepts, including the rules of probability, conditional probability, and Bayes' theorem. These concepts are widely used in AI and machine learning, particularly in classification problems, predictive modeling, and decision-making processes.



Basic Probability Concepts

Definition of Probability

Probability measures the likelihood of an event occurring. It ranges from 0 to 1, where 0 indicates an impossible event and 1 indicates a certain event.

Example

The probability of getting heads in a fair coin toss is 0.5.

Random Events and Outcomes

Random Event

- An event whose occurrence cannot be predicted with certainty. Examples include rolling a die, flipping a coin, or picking a card from a deck.

Outcomes

- The possible results of a random event. For a coin toss, the possible outcomes are heads or tails. For rolling a six-sided die, the outcomes are 1, 2, 3, 4, 5, and 6.

Basic Probability Concepts

Probability Formula

For an event A, the probability $P(A)$ is calculated as:

$$P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$$

Example

The probability of drawing a heart from a standard deck of cards:

$$P(\text{heart}) = \frac{13}{52} = \frac{1}{4} = 0.25$$

Probability Rules

Addition Rule (For Mutually Exclusive Events)

If two events A and B cannot happen at the same time (i.e., they are mutually exclusive), the probability that either A or B will occur is:

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

Example

The probability of drawing either a king or a queen from a deck of cards:

$$P(\text{king or queen}) = P(\text{king}) + P(\text{queen}) = \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13}$$

Probability Rules

Multiplication Rule (For Independent Events)

If two events A and B are independent (the occurrence of one does not affect the occurrence of the other), the probability that both A and B will occur is:

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

Example

The probability of getting heads when flipping a coin and rolling a 4 on a six-sided die:

$$P(\text{heads and } 4) = P(\text{heads}) \times P(4) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$$

Conditional Probability

Conditional probability measures the probability of an event A occurring given that event B has occurred. It is denoted by $P(A|B)$.

Formula:
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Example

In a deck of cards, what is the probability of drawing a queen given that you have already drawn a king?

- Without replacement, after drawing a king, there are 51 cards left. There are still 4 queens.

$$P(\text{queen}|\text{king}) = \frac{4}{51}$$

Bayes' Theorem

Introduction to Bayes' Theorem: Bayes' theorem allows us to update our probability estimates based on new **evidence** or **information**. It is particularly useful in AI for making predictions based on **prior knowledge** and new data.

Bayes' Theorem Formula:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Where:

- $P(A|B)$ is the posterior probability: the probability of event A given B has occurred.
- $P(B|A)$ is the likelihood: the probability of event B given A is true.
- $P(A)$ is the prior probability: the probability of event A before observing B .
- $P(B)$ is the marginal likelihood: the total probability of event B .

Bayes' Theorem

Example

Medical Diagnosis

- Suppose a certain disease affects 1% of the population ($P(\text{Disease})=0.01$).
- A test correctly identifies the disease 99% of the time ($P(\text{Positive Test}|\text{Disease})=0.99$).
- However, the test also produces false positives 5% of the time ($P(\text{Positive Test}|\text{No Disease})=0.05$).

Question

What is the probability that a person has the disease given they tested positive?

Bayes' Theorem

Example

Using Bayes' theorem:

$$P(\text{Disease}|\text{Positive Test}) = \frac{P(\text{Positive Test}|\text{Disease}) \times P(\text{Disease})}{P(\text{Positive Test})}$$

Calculate $P(\text{Positive Test})$:

$$P(\text{Positive Test}) = P(\text{Positive Test}|\text{Disease}) \times P(\text{Disease}) + P(\text{Positive Test}|\text{No Disease}) \times P(\text{No Disease})$$

$$P(\text{Positive Test}) = (0.99 \times 0.01) + (0.05 \times 0.99) = 0.0099 + 0.0495 = 0.0594$$

$$\text{Apply Bayes' Theorem: } P(\text{Disease}|\text{Positive Test}) = \frac{0.99 \times 0.01}{0.0594} \approx 0.1667$$

This means there is approximately a 16.67% chance that the person actually has the disease, despite testing positive. This highlights how Bayes' theorem can provide more accurate probability estimates when additional information is available.

Introduction to Inferential Statistics (30 mins)

Introduction to Inferential Statistics

What is Inferential Statistics?

Inferential statistics involve using data from a sample to make generalizations about a larger population. **Unlike** descriptive statistics, which summarize data, inferential statistics allow us to draw conclusions and make predictions.

Purpose in AI and Data Science

- Helps estimate population parameters (e.g., the mean, proportion).
- Allows for hypothesis testing to determine if observed patterns are likely to be genuine or if they occurred by chance.
- Facilitates predictions and decision-making based on sample data.

Key Concepts of Inferential Statistics

Sampling Distributions

A sampling distribution is the probability distribution of a given statistic (e.g., mean, proportion) based on a random sample.

Importance

- Sampling distributions help understand how the sample statistic (like the sample mean) varies from sample to sample.
- They form the basis for making inferences about the population.

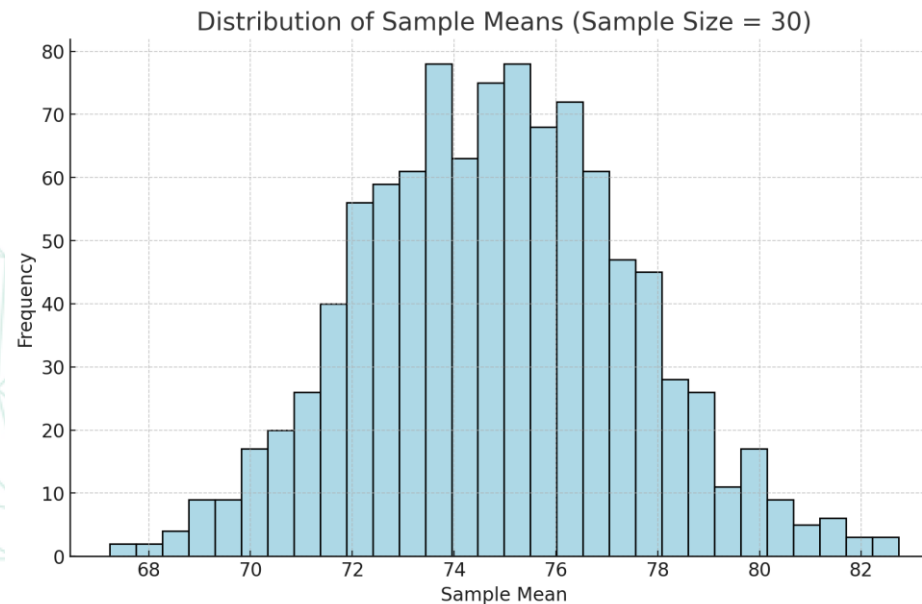
Example

If we take multiple random samples of students' test scores and calculate the mean for each sample, the distribution of those sample means would form a sampling distribution. This distribution tends to be normal, especially as the sample size increases, according to the Central Limit Theorem.

Key Concepts of Inferential Statistics

Example

If we take multiple random samples of students' test scores and calculate the mean for each sample, the distribution of those sample means would form a sampling distribution. This distribution tends to be normal, especially as the sample size increases, according to the Central Limit Theorem.



Key Concepts of Inferential Statistics

Central Limit Theorem (CLT)

The CLT states that the sampling distribution of the sample mean will be approximately normally distributed, regardless of the population's distribution, provided the sample size is sufficiently large (usually $n > 30$).

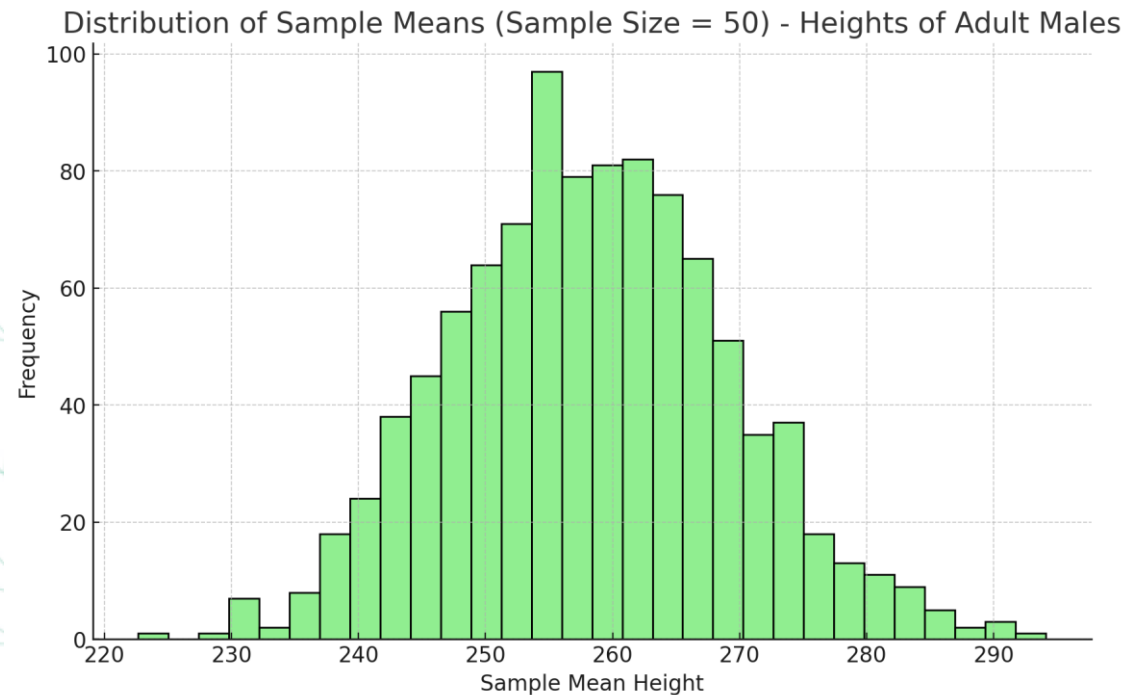
Significance

The CLT enables us to use the normal distribution to make inferences about population parameters even when the population distribution is not normal.

Key Concepts of Inferential Statistics

Example

Suppose we are analyzing the average height of adult males. Even if the population distribution of heights is skewed, the mean of sufficiently large samples will follow a normal distribution.



Hypothesis Testing

Hypothesis testing is a statistical method used to make inferences about population parameters based on sample data. It helps us decide whether there is enough evidence to reject a null hypothesis.

Key Terms

- **Null Hypothesis (H_0):** A statement that there is no effect or no difference. It serves as the default assumption. For example, "The mean score of students is equal to 75."
- **Alternative Hypothesis (H_1 or H_a):** The statement that contradicts the null hypothesis. For example, "The mean score of students is not equal to 75."
- **P-value:** The probability of observing the sample data, or something more extreme, assuming the null hypothesis is true. A low p-value (typically < 0.05) suggests that the observed data is unlikely under the null hypothesis, leading to its rejection.
- **Significance Level (α):** A threshold chosen before conducting the test, often set at 0.05, indicating a 5% risk of concluding that a difference exists when there is no actual difference.

Hypothesis Testing

Steps in Hypothesis Testing

1. **State the Hypotheses:** Define the null and alternative hypotheses.
2. **Select a Significance Level (α):** Commonly set at 0.05.
3. **Collect Data and Calculate a Test Statistic:** Based on the sample data, calculate a statistic (e.g., t-statistic, z-statistic) that measures the difference.
4. **Find the P-value:** Compare the test statistic to the theoretical distribution to find the p-value.
5. **Make a Decision:**
 - If the p-value $< \alpha$, reject the null hypothesis (evidence suggests an effect).
 - If the p-value $\geq \alpha$, do not reject the null hypothesis (insufficient evidence to suggest an effect).

Hypothesis Testing

Example

A company claims that their employees work an average of 8 hours per day. A random sample of 30 employees shows an average of 7.5 hours with a standard deviation of 1 hour. We could conduct a hypothesis test to determine if this sample provides enough evidence to reject the company's claim.

Confidence Intervals

A confidence interval provides a range of values that is likely to contain the population parameter (e.g., mean) with a certain level of confidence (e.g., 95%).

Formula : $\text{Confidence Interval} = \bar{x} \pm Z \left(\frac{\sigma}{\sqrt{n}} \right)$

- \bar{x} = Sample mean
- Z = Z-value (from standard normal distribution for a given confidence level)
- σ = Population standard deviation (or sample standard deviation if population standard deviation is unknown)
- n = Sample size

Interpretation

A 95% confidence interval means that if we took 100 different samples and computed a confidence interval for each sample, approximately 95 of the intervals would contain the true population mean.

Confidence Intervals

Example

If a sample mean score on a test is 70 with a standard deviation of 5 and sample size of 30, a 95% confidence interval for the mean score might be calculated. This interval provides a range within which we expect the true mean score for the entire population to fall.