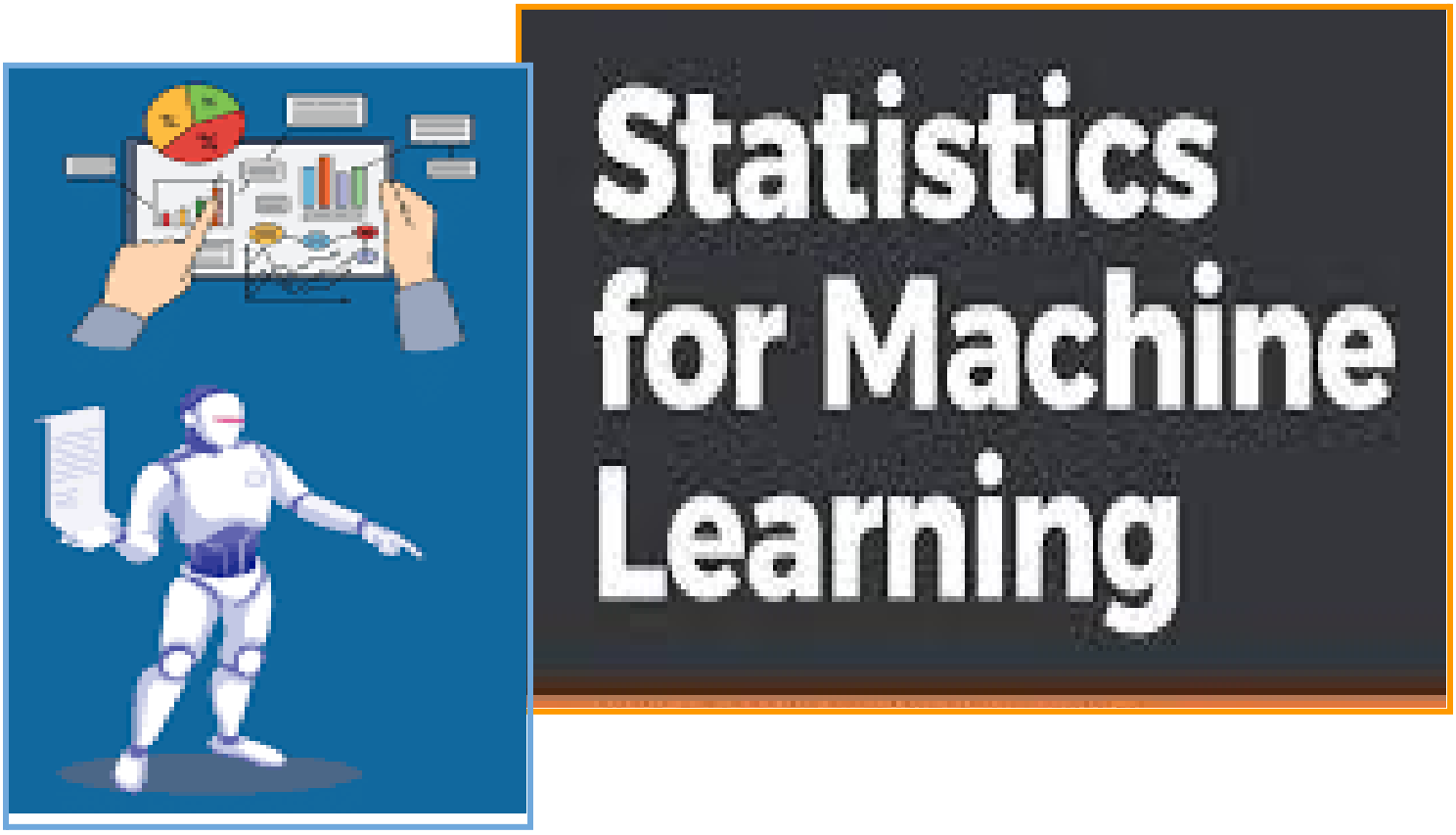
Applied Statistics Interview Grind Project



Sowmya Bhupatiraju

In today's highly competitive job market, strong conceptual understanding and clear communication of problem-solving approaches are essential for success in technical interviews. Given this, the project aims to enhance the ability to clearly communicate technical concepts, both in written form and through verbal explanation.

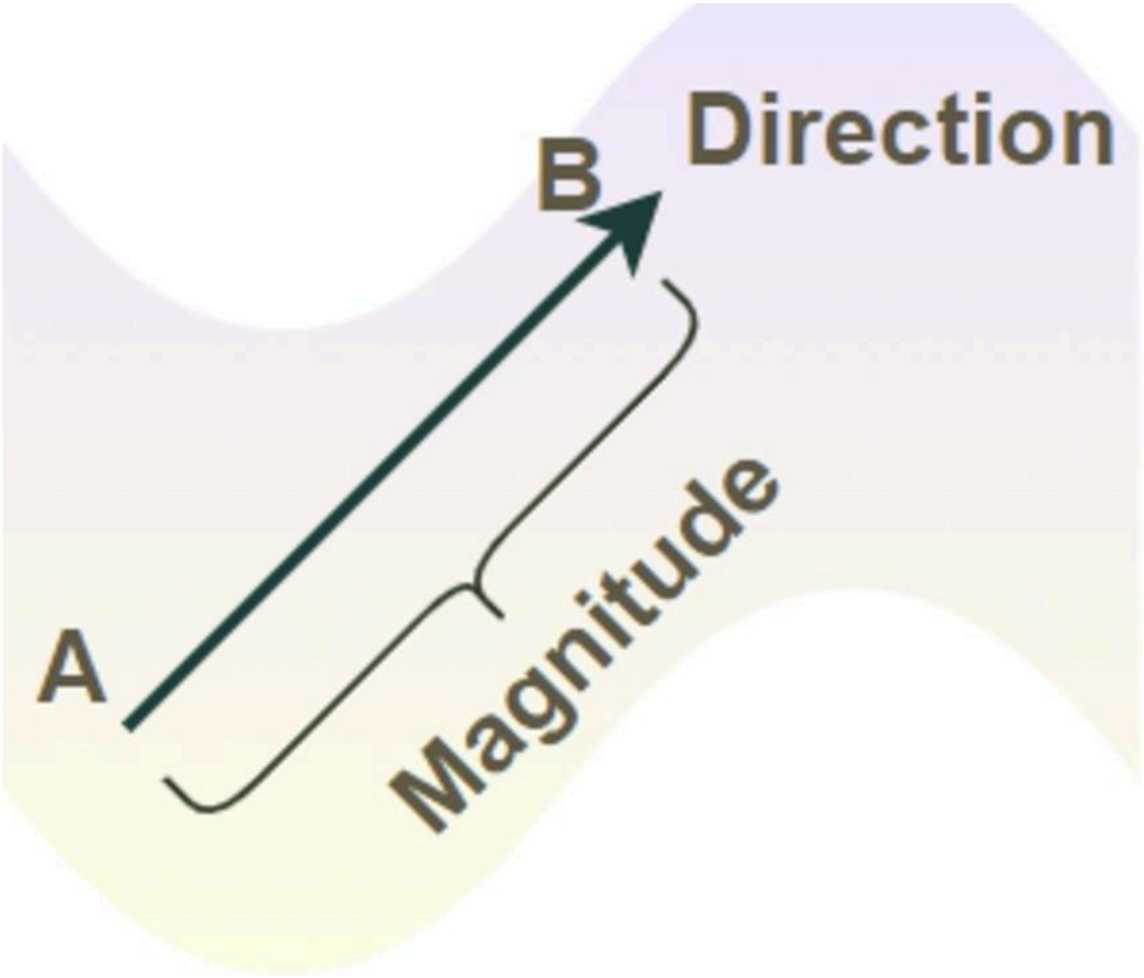
As part of this assignment, I have selected 50 interview questions from a given set of 80 and provided concise, well-structured, and solution-oriented answers. These solutions demonstrate my understanding of core concepts across relevant technical domains while emphasizing clarity, correctness, and real-world applicability.

Overall, this project reflects my preparation for technical interviews, my analytical thinking capabilities, and my commitment to continuous learning and effective knowledge sharing.

1. What is a vector in mathematics?(1)

A vector is a mathematical quantity with both magnitude and direction.

Magnitude defines the size of the vector. It is represented by a line with an arrow, where the length of the line is the magnitude of the vector and the arrow shows the direction.



We can calculate the magnitude of the vector by taking the square root of the sum of the squares of each component in the x and y directions.

$$|A| = \sqrt{a^2 + b^2}$$

The magnitude of a vector is a scalar value.

***Example**: $\bar{A} = 2i + 3\hat{j}$*
Magnitude $|A| = \sqrt{4 + 9} = \sqrt{13} = 3.61$

If two vectors in the 2-D plane intersect each other then the angle between them can easily be calculated using the formula

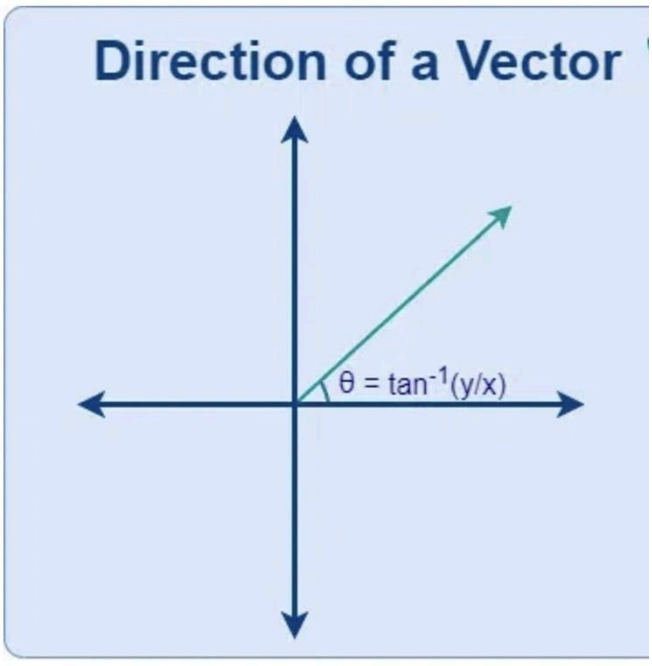
$$\theta = \tan^{-1}(y/x)$$

$$= \tan^{-1}(3/2)$$

$$\approx 0.983 \text{ radians}$$

where:

- *θ is the angle between the vector and the positive x-axis.*
- *$\tan^{-1}$ represents the inverse tangent function (also called arctan).*



Direction of a Vector

Note:

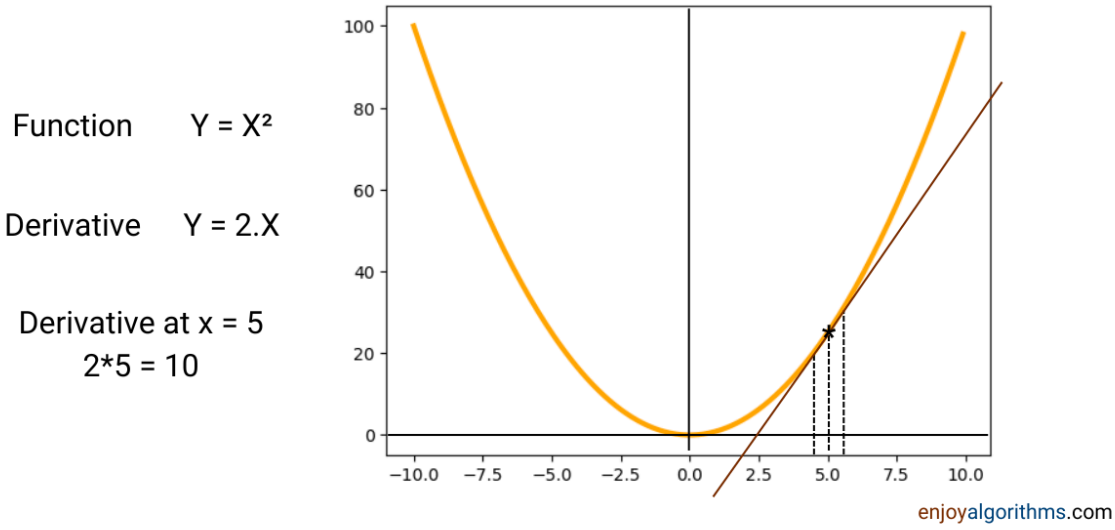
- *This formula gives the angle in radians.*
- *You may need to adjust the angle based on the quadrant in which the vector lies to get the correct direction.*

Example: If a vector v = (3, 4), then:

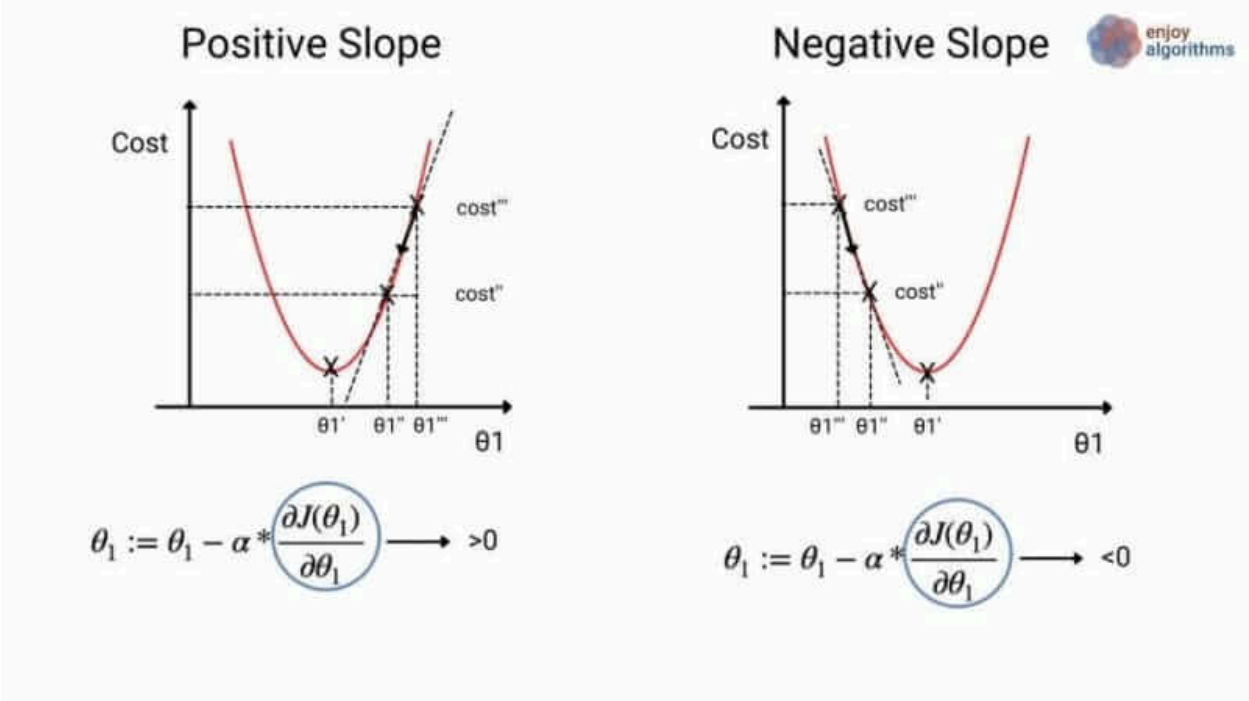
=> θ = arctan(4/3) ≈ 53.13 degrees

2. Why do we need derivatives in Machine Learning?

*Derivates of a function at any point describe or the slope of the tangent line to its graph at a point or measures the instantaneous rate of change of a function, . For example, the image below shows the derivate of f(x) = x², at x = 5. From the derivative theory, differentiating x² with respect to x will produce 2*x. So, at x=5, the derivative of f(x) = x² will be 2*5 = 10, a positive value. The sign of this derivative at a certain point conveys critical information about the curve at that point.*



*In machine Learning, f(x) is the cost function, x is the parameter and the goal is to achieve that value of x for which f(x) is minimum. Parameter update happens in this way: θ:= θ—α*d(J(θ))/d(θ). Here, α is known as the learning parameter, and one can find the requirement/effects of α. For example, in the image below, θ1' is the desired parametric value machines want to learn, and the updation equation is: θ1:= θ1—α*d(J(θ1))/d(θ1).*



In starting, the machine will randomly pick any 'θ1', and only three possible scenarios can be expected:

θ1 > θ1': The optimization algorithm will ask ML algorithms to reduce the value for θ1 so that θ1 → θ1'.
θ1 < θ1': The optimization algorithm will ask ML algorithms to increase the value for θ1 so that θ1 → θ1'.
θ1 = θ1': No update in the parameter value is required. ML process will be completed.

The update parameter uses a negative (-) of the derivative. Hence, in the case of a positive slope, the value for that parameter will decrease; if the slope is negative, the value for the parameter will increase.

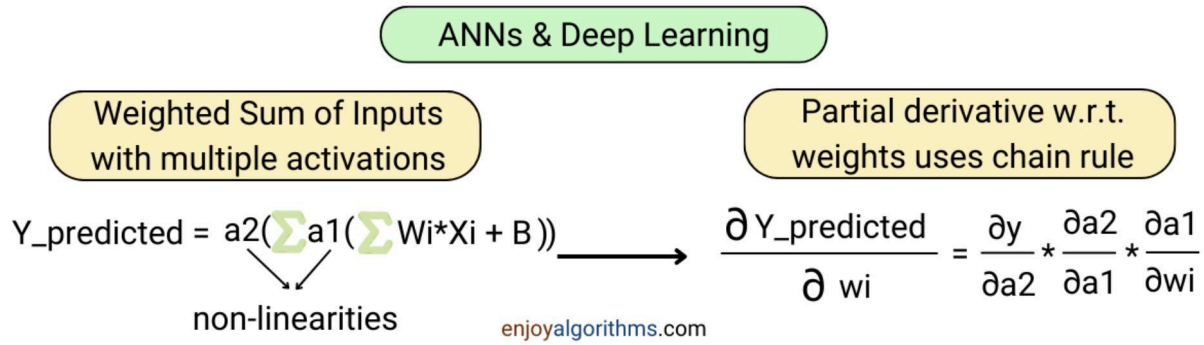
In most machine learning use cases, multiple learnable parameters are involved in the learning process. Machines intended to find the optimum value for all those parameters. If any ML model involves "n" parameters, all these parameters will somehow affect the cost function. So, we can represent the cost function as J(θ1, θ2, ..., θn). All these parameters need to be updated in the right direction to find the minimum value for J.

So, derivatives are the mathematical tools used to minimize error and optimize performance. The primary uses of derivatives are in calculating the gradients in Gradient Descent Algorithm and other optimization algorithms like Stocahastic Gradient Descent, ADAM etc

But if we calculate the derivative of this cost function (which depends on' n' variables), all constituent variables will start affecting the update process for one variable. ML algorithms will find it difficult to sense whether to increase or decrease the value. That's why the concept of partial derivative comes into the picture.

This is where we need partial derivatives in Machine Learning.

While calculating the partial derivative of a function with respect to any one parameter out of multiple parameters, we treat that parameter as the only variable affecting that function at that time. Rest all parameters are considered to be constant for that function. This helps to focus on one parameter during the updation process, and that's why we use partial derivatives in machine learning.



3. How are partial derivatives used in machine learning?

While calculating the partial derivative of a function with respect to any one parameter out of multiple parameters, we treat that parameter as the only variable affecting that function at that

time. Rest all parameters are considered to be constant for that function and hence partial derivatives are fundamental in machine learning for optimizing models by measuring how a cost function changes with respect to individual parameters (weights and biases). They enable gradient-based optimization, such as gradient descent and backpropagation, allowing models to update parameters, minimize error, and improve performance by calculating the gradient, which indicates the direction of steepest descent.

Partial derivatives are fundamental in machine learning for ***optimizing models by measuring how a cost function changes with respect to individual parameters (weights and biases)****. They enable gradient-based optimization, such as gradient descent and backpropagation, allowing models to update parameters, minimize error, and improve performance by calculating the gradient, which indicates the direction of steepest descent.*

Key Uses of Partial Derivatives in Machine Learning:

- ***Gradient Descent Optimization:*** *To minimize the error (loss) of a model, partial derivatives are calculated for each parameter to determine the direction of the steepest ascent/descent. This allows the algorithm to update parameters (e.g., $w = w - \alpha \frac{\partial J}{\partial w}$) iteratively, reducing the error.*
- ***Backpropagation in Neural Networks:*** *During training, partial derivatives are used to compute the gradient of the loss function with respect to each weight in the network. This involves applying the chain rule to propagate errors backward through layers.*
- ***Chain Rule Application:*** *Since neural networks are complex composite functions (nested functions), partial derivatives are combined via the chain rule to calculate the contribution of each neuron to the final error.*
- ***Sensitivity Analysis and Feature Importance:*** *Partial derivatives help determine how sensitive the model's output is to small changes in specific input features, aiding in understanding feature importance.*
- ***Regularization Techniques:*** *They are used to calculate penalty terms in loss functions (like L1 and L2 regularization) that prevent overfitting.*

By treating other variables as constants, partial derivatives focus on how a single parameter contributes to the overall loss, allowing for precise, iterative model improvements.

4. What is an eigenvector in linear algebra?(10)

Eigenvalues and Eigenvectors *are the scalar and vector quantities associated with matrices used for linear transformations. The vector that only changes by a scalar factor after applying a transformation is called an eigenvector, and the scalar value attached to the eigenvector is called the eigenvalue.*

The equation for eigenvalue is given by

$$Av = \lambda v$$

Where,

- *A is the matrix,*
- *v is associated eigenvector, and*
- *λ is scalar eigenvalue.*

Eigenvectors *are the directions that remain unchanged(or reversed) during a transformation, even if they get longer or shorter.* ***Eigenvalues*** *are the numbers that indicate how much something stretches or shrinks during that transformation. Eigenvalues and eigenvectors are important concepts in linear algebra and have various applications in data science and machine learning - to transform and reduce the dimensionality of data in various algorithms like PCA, SVD, and others.*

Example : Find the eigenvectors of the matrix A = [[5, 0], [0, 5]]

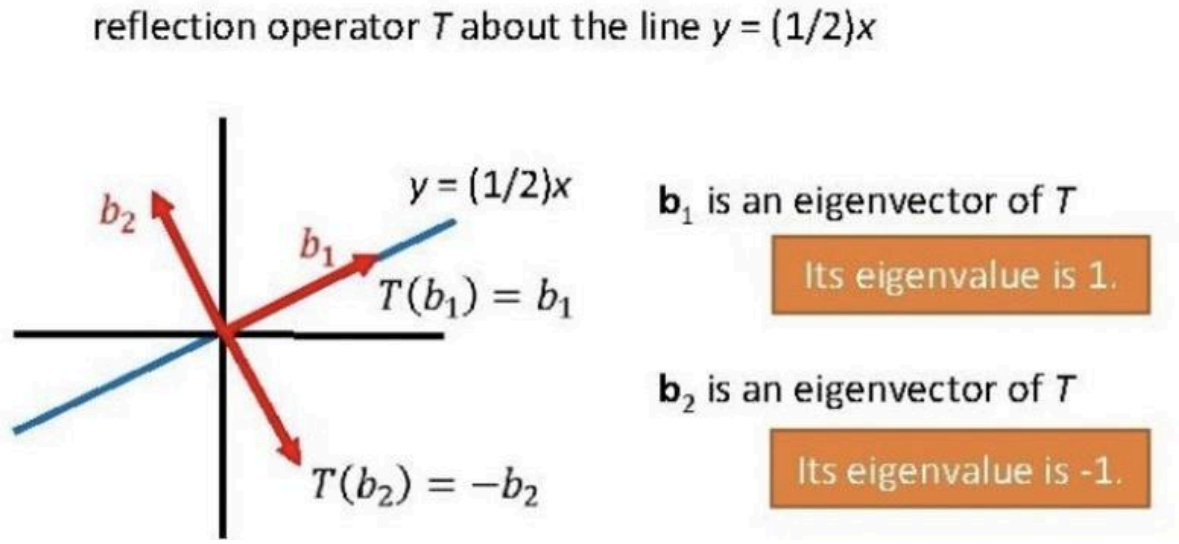
1. ***Find the eigenvalues:***
    - *The characteristic equation is |A - λI| = 0, where λ is the eigenvalue and I is the identity matrix.*
    - *For matrix A, the characteristic equation is: |[[5, 0], [0, 5]] - λ[[1, 0], [0, 1]]| = 0 |[[5-λ, 0], [0, 5-λ]]| = 0 (5 - λ)(5 - λ) = 0*
    - *Solving for λ, we get the eigenvalues: λ1 = 5, λ2 = 5*
2. ***Find the eigenvectors:***
    - ***For eigenvalue λ1 = 5:***
        - *Solve the equation (A - λ1I)v1 = 0, where v1 is the eigenvector corresponding to λ1.*
        - *[[5-5, 0], [0, 5-5]] * [x1, y1] = [0, 0]*
        - *[[0, 0], [0, 0]] * [x1, y1] = [0, 0]*
        - *This gives us the equation 0x1 + 0y1 = 0, which is satisfied by any vector [x1, y1].*
        - *Therefore, the eigenvectors corresponding to λ1 = 5 are of the form [x1, 0] where x1 is any non-zero scalar. For simplicity, we can choose v1 = [1, 0].*
    - ***For eigenvalue λ2 = 5:***
        - *Solve the equation (A - λ2I)v2 = 0, where v2 is the eigenvector corresponding to λ2.*
        - *[[5-5, 0], [0, 5-5]] * [x2, y2] = [0, 0]*
        - *[[0, 0], [0, 0]] * [x2, y2] = [0, 0]*

- *This gives us the equation 0x2 + 0y2 = 0, which is satisfied by any vector [x2, y2].*
- *Therefore, the eigenvectors corresponding to λ2 = 5 are of the form [0, y2] where y2 is any non-zero scalar. For simplicity, we can choose v2 = [0, 1].*

Therefore, the eigenvectors of the matrix A are:

- *v1 = [1, 0]*
- *v2 = [0, 1]*

Note: *Since the matrix A is a diagonal matrix, the eigenvectors are simply the standard basis vectors.*



5. What is the gradient in machine learning?(11)

Gradient is simply a partial derivative of a function with respect to its inputs..

You can also think of a gradient as the slope of a function or represents rate of change of a function with respect to its input parameters and is used to determine the direction in which to update the parameters.

In machine learning, the concept of gradient vector is used for optimization purposes - to reduce the error between predicted and actual results by iteratively adjusting a model's

parameters to minimize (or the cost function) a certain objective function - (or the cost function) in the opposite direction of the gradient.

To understand how gradients work, let's consider a ***simple example.*** *Suppose we have a function f(x) = x². The gradient of this function with respect to x is dx / df = 2x. This gradient tells us how f changes as we vary x. For instance, when x = 2, the gradient is 2 × 2 = 4, indicating that f increases at a rate of 4 units for every unit increase in x.*

Gradient descent algorithm

repeat until convergence {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

(for $j = 1$ and $j = 0$)

}

Linear Regression Model

$$h_\theta(x) = \theta_0 + \theta_1 x$$

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

But the update parameter uses a negative (-) of the derivative as given above. Hence, in the case of a positive slope, the value for that parameter will decrease; if the slope is negative, the value for the parameter will increase, thereby decreasing the cost function - one step at a time to reach global minima.

6. What is backpropagation in machine learning?

Backpropagation *is a powerful algorithm in deep learning, primarily used to train artificial neural networks, particularly* ***feed-forward networks****. It works iteratively, minimizing the cost function by adjusting weights and biases.*

In each epoch, the model adapts these parameters, reducing loss by following the error gradient. Backpropagation often utilizes optimization algorithms like ***gradient descent*** *or* ***stochastic gradient descent****. The algorithm computes the gradient using the chain rule from calculus, allowing it to effectively navigate complex layers in the neural network to minimize the cost function*

In the backward pass, the error (the difference between the predicted and actual output) is propagated back through the network to adjust the weights and biases. One common method for error calculation is the ***Mean Squared Error (MSE)****, given by:*

$MSE = (1/n) * \Sigma(predicted\_output - actual\_output)^2$

Once the error is calculated, the network adjusts weights using ***gradients****, which are computed with the chain rule. These gradients indicate how much each weight and bias should be adjusted to minimize the error in the next iteration. The backward pass continues layer by layer, ensuring that the network learns and improves its performance. The activation function, through its derivative, plays a crucial role in computing these gradients during backpropagation.*

Example of Backpropagation in Machine Learning

Let's walk through an example of backpropagation in machine learning. Assume the neurons use the sigmoid activation function for the forward and backward pass. The target output is 0.5, and the learning rate is 1.

1. Calculating Gradients

The change in each weight is calculated as:

$$\Delta wij = \eta * \delta j * Oi$$

Where:

- δj *- the error term for each unit.*

- *η - learning rate.*
- *Oi - output of the previous unit.*

2. Output Unit Error

For O3 :

Given:

- ***Output Unit Error (δ5):*** *δ5 = y5 * (1 - y5) * (target - y5)*
- *δ5 = 0.67 * (1 - 0.67) * (-0.17) = -0.0376*
- ***Hidden Unit Errors (δ3 and δ4):*** *δ3 = y3 * (1 - y3) * (w1,3 * δ5)*

*=> δ3 = 0.56 * (1 - 0.56) * (0.3 * -0.0376) = -0.0027*

- *δ4 = y4 * (1 - y4) * (w2,3 * δ5)*

*=> δ4 = 0.59 * (1 - 0.59) * (0.9 * -0.0376) = -0.0819*

- ***Learning Rate (η):*** *Assuming a learning rate of 1 for simplicity.*

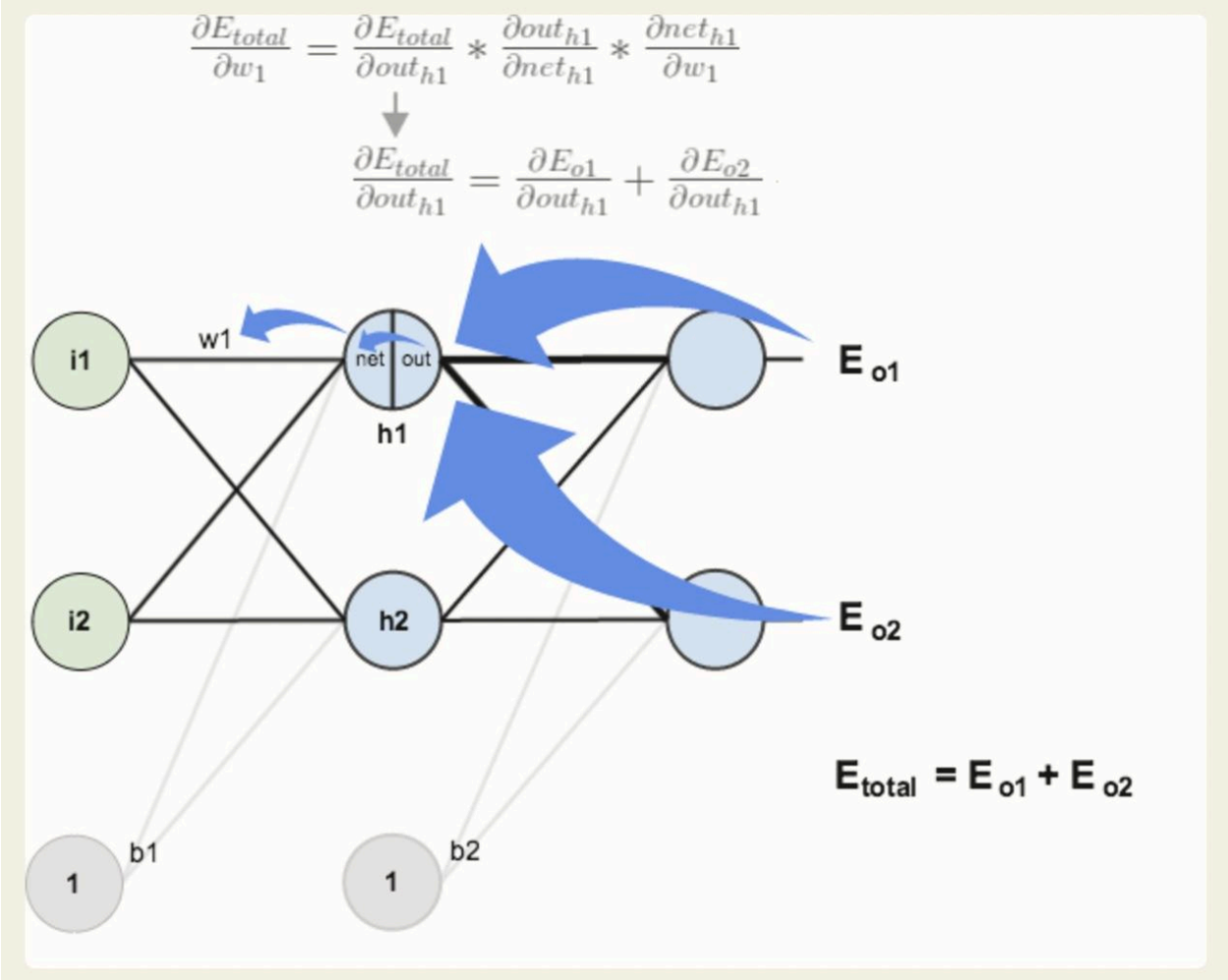
Weight Updates:

- ***For the weight w2,3:*** *Δw2,3 = η * δ5 * y4*

*=> 1 * (-0.0376) * 0.59*
=> -0.022184

- *New w2,3 = w2,3 (old) + Δw2,3 = 0.9 + (-0.022184) = 0.877816*

Note: ***Backpropagation*** *is an iterative process. This example demonstrates a single weight update step. In practice, multiple iterations would be performed to refine the weights and minimize the overall error.*



7. What is probability theory?

Probability theory *is an advanced branch of mathematics that deals with* ***measuring the likelihood*** *of events occurring. It provides tools to analyze situations involving* ***uncertainty*** *and helps in determining how likely certain outcomes are. This theory uses the concepts of random variables, sample space, probability distributions, and more to determine the outcome of any situation.*

Probability theory studies random events and tells us about their occurrence.

The three main approaches for studying probability theory are:

- *Theoretical Probability*
- *Experimental Probability*
- *Subjective Probability*

Theoretical Probability: *Ex - The Probability of the occurrence of a Tail on tossing a coin is P(T) = 1/2*

Experimental Probability: Example-*If we tossed a coin 10 times and recorded heads for 4 times and a tail 6 times then the Probability of occurrence of Head on tossing a coin: P(H) = 4/10*

Similarly, the Probability of Occurrence of Tails on tossing a coin: P(T) = 6/10

Subjective Probability: Example: *A cricket enthusiast might assign a 70% probability to a team's victory based on their understanding of the team's recent form, the opponent's strengths and weaknesses, and other relevant factors.*

The different types of probabilities - ***unconditional probabilities, joint probabilities, and conditional probabilities.***
Some of the important uses of probability theory are:

- *Probability theory is used to predict the performance of stocks and bonds.*
- *In casinos and gambling probability theory is used to find the chances of winning.*
- *Probability theory is used in weather forecasting.*
- *Probability theory is used in Risk mitigation.*

8. What are the primary components of probability theory?
Random Experiment: *In probability theory, any event that can be repeated multiple times and its outcome is not hampered by its repetition is called a Random Experiment. Tossing a coin, rolling dice, etc. are random experiments.*

Sample Space: *The set of all possible outcomes for any random experiment is called sample space. For example, throwing dice results in six outcomes, which are 1, 2, 3, 4, 5, and 6. Thus, its sample space is (1, 2, 3, 4, 5, 6)*

Event: *The outcome of any experiment is called an event. Various types of events used in probability theory are Independent Events, Dependent Events, Mutually Exclusive Events and Equally likely Events.*

Random Variable: *A variable that can assume the value of all possible outcomes of an experiment is called a random variable in Probability Theory. Random variables in probability theory are of two types namely Continuous Random Variable and Discrete Random Variable.*

Probability Distributions: *A function that describes the probability of a random variable taking on different values.The properties of a random variable can be described by its probability distribution, which specifies the probabilities of each possible value of the variable. The probability distribution of a discrete random variable can be represented by a* ***probability mass function (PMF-Examples of discrete probability distributions include the binomial distribution, the Poisson distribution, and the geometric distribution.)****, while the probability distribution of a continuous random variable can be represented by a* ***probability density function (PDF-Some examples of continuous probability distributions are Normal Distribution, Uniform Distribution, Exponential Distribution, Beta Distribution, Gamma Distribution, and Weibull Distribution).***

In Python, we can use libraries like NumPy and SciPy to work with random variables and their properties.

Probability Theory Formulas: *Some of the formulas that are commonly used in probability theory are discussed below:*

- ***Theoretical Probability Formula:*** *(Number of Favourable Outcomes) / (Number of Total Outcomes)*
- ***Empirical Probability Formula:*** *(Number of times event A happened) / (Total number of trials)*
- ***Addition Rule of Probability:*** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- ***Complementary Rule of Probability:*** $P(A') = 1 - P(A)$
- ***Independent Events:*** $P(A \cap B) = P(A) \cdot P(B)$
- ***Conditional Probability:*** $P(A \mid B) = P(A \cap B) / P(B)$
- ***Bayes' Theorem:*** $P(A \mid B) = P(B \mid A) \cdot P(A) / P(B)$
- ***Expected Value:*** *The average value of a random variable.*
- ***Variance and Standard Deviation:*** *Measures of the spread of a probability distribution.*

The ***law of large numbers and central limit theorem*** *are fundamental theorems in probability theory that help to understand the behavior of sample means and their distribution as sample size increases*

Applications in Statistics:

- ***Descriptive Statistics****: Probability theory helps in understanding and interpreting data summaries and distributions.*
- ***Inferential Statistics****: This forms the basis for making inferences about populations from samples, including hypothesis testing and the construction of confidence intervals.*

- ***Regression Analysis****: Probability distributions of errors are used to estimate the relationships between variables.*
- ***Bayesian Statistics****: Uses probability to represent uncertainty about the parameters of interest and updates this uncertainty as more data becomes available.*

9. What is conditional probability, and how is it calculated?(17)

Conditional probability is the probability of an event A given that another event B has already occurred. It is denoted by P(A|B), which means the probability of A given B. The formula for conditional probability is:

P(A|B) = P(A ∩ B) / P(B)

Where

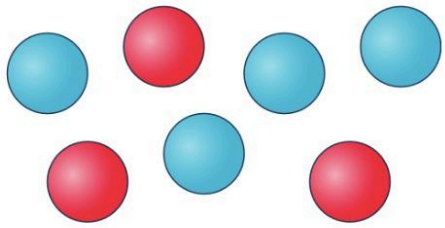
P(A ∩ B) *is the probability of both A and B occurring, and*

P(B) *is the probability of B occurring.*

Example*: Suppose you have two decks of cards, one red and one blue, with 52 cards each. You draw a card from the red deck and observe that it is a heart. What is the probability that the card is an ace?*

Here, A is the event of drawing an ace, and B is the event of drawing a heart. The probability of drawing an ace and a heart is P(A ∩ B) = 1/52, and the probability of drawing a heart is P(B) = 13/52. Using the formula for conditional probability, we get P(A|B) = P(A ∩ B) / P(B) = (1/52) / (13/52) = 1/13, which is the probability of drawing an ace given that the card is a heart.

You have a bag containing 3 red marbles and 4 blue marbles.

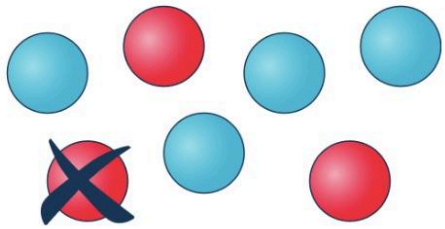


The probabilities of picking each color marble are,

$$P(red\ marble) = \frac{3}{7}$$

$$P(blue\ marble) = \frac{4}{7}$$

If you picked out one red marble, there would be 2 red marbles and 4 blue marbles left.



The probabilities would now be,

$$P(red\ marble) = \frac{2}{6}$$

$$P(blue\ marble) = \frac{4}{6}$$

10. What is Bayes theorem, and how is it used?(18)

Bayes' theorem is a fundamental concept in probability theory, and it is used to calculate conditional probabilities. It states that the probability of an event A given that another event B has occurred can be calculated as:

P(A|B) = P(B|A) * P(A) / P(B)

Where

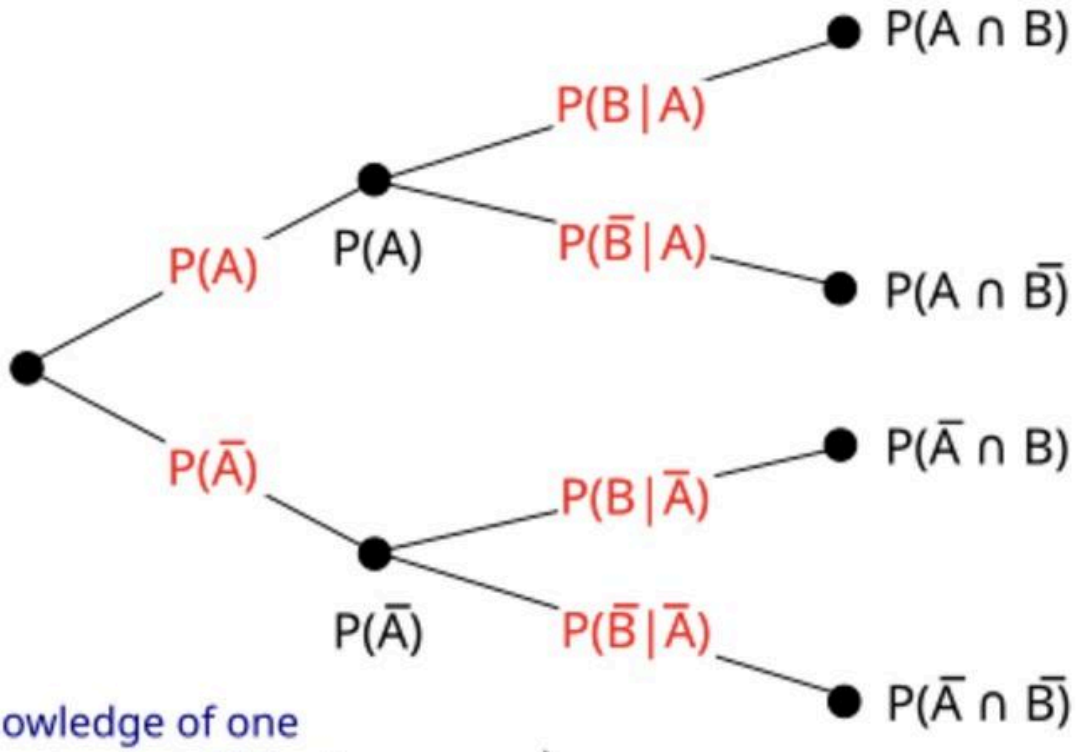
P(B\|A) *is the probability of B given A has occurred,*

P(A) *is the prior probability of A, and*

P(B) *is the prior probability of B.*

Bayes Theorem *is the extension of Conditional probability and talks about the relation of the conditional probability of two random events and their marginal probability.*
Bayes' theorem can be used to update our beliefs about the probability of an event as new information becomes available.Here are some important applications of Bayes' rule in AI.

1. ***Bayesian Inference:*** *In Bayesian statistics, the Bayes' rule is used to update the probability distribution over a set of parameters or hypotheses using observed data. This is especially important for machine learning tasks like parameter estimation in Bayesian networks, hidden Markov models, and probabilistic graphical models.*
2. ***Naive Bayes Classification:*** *In the field of natural language processing and text classification, the Naive Bayes classifier is widely used. It uses Bayes' theorem to calculate the likelihood that a document belongs to a specific category based on the words it contains. Despite its "naive" assumption of feature independence, it works surprisingly well in practice.*
3. ***Bayesian Networks****: Bayesian networks are graphical models that use Bayes' theorem to represent and predict probabilistic relationships between variables. They are used in a variety of AI applications, such as medical diagnosis, fault detection, and decision support systems.*
4. ***Reinforcement Learning****: Bayes' rule can be used to model the environment in a probabilistic manner. Bayesian reinforcement learning methods can help agents estimate and update their beliefs about state transitions and rewards, allowing them to make more informed decisions.*
5. ***Personalization:*** *In recommendation systems, Bayes' theorem can be used to update user preferences and provide personalized recommendations. By constantly updating a user's preferences based on their interactions, the system can recommend more relevant content.*
6. ***Robotics and Sensor Fusion:*** *In robotics, the Bayes' rule is used to combine sensors. It uses data from multiple sensors to estimate the state of a robot or its environment. This is necessary for tasks like localization and mapping.*
7. *Other uses include* ***Spam Email Filtering, Bayesian Optimization, Anomaly Detection, Medical Diagnosis***

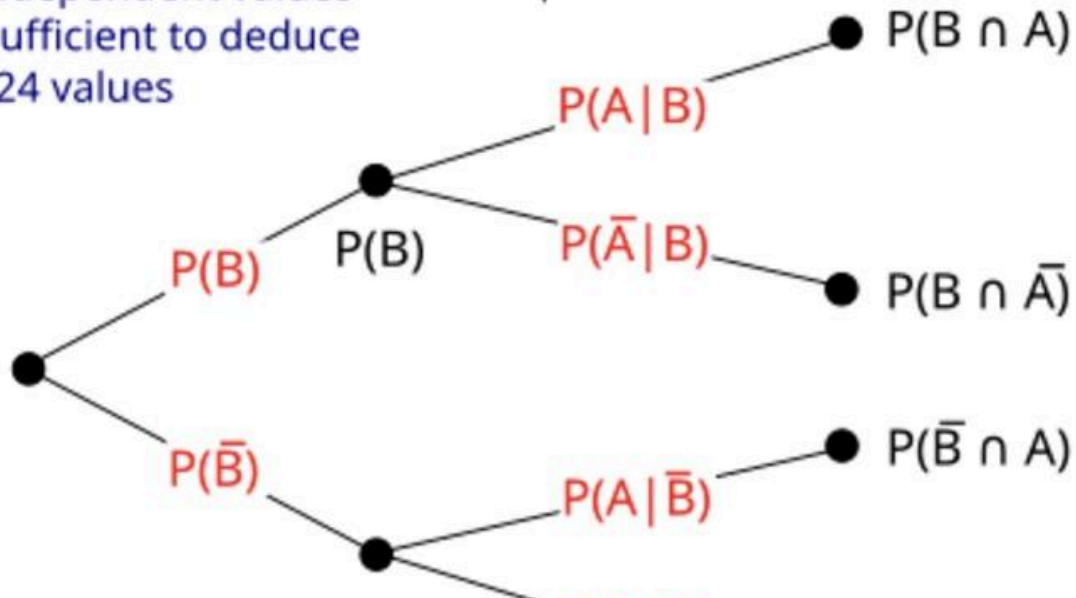


Knowledge of one diagram is sufficient to deduce the other

Use Bayes' Theorem to convert between diagrams

$$P(\alpha|\beta)\,P(\beta) = P(\alpha \cap \beta) = P(\beta|\alpha)\,P(\alpha)$$

Knowledge of any 3 independent values is sufficient to deduce all 24 values



11. What is a random variable, and how is it different from a regular variable?(19)

A variable and a random variable are both concepts used in mathematics and statistics, but they have distinct meanings

Regular Variable: *Represents a fixed or known quantity. Its value is deterministic and predictable. For example, in the equation x + 5 = 10, the value of x is a regular variable and can be determined definitively.*

Random variable *is a specific type of variable that is used in probability and statistics to represent the outcomes of a random phenomenon.A random variable is a function that maps each outcome of a random process to a numerical value. For example, if we roll a die, the random variable can be defined as the number that appears on the top face.*

- *Random variables can be classified into two types: Discrete and continuous. A discrete random variable can take on a countable number of values, while a continuous random variable can take on any value in a continuous range. Examples of discrete random variables include the number of heads in multiple coin flips, while an example of a continuous random variable is the height of a randomly selected person.*
- *The properties of a random variable can be described by its probability distribution, which specifies the probabilities of each possible value of the variable. The probability distribution of a discrete random variable can be represented by a probability mass function (PMF), while the probability distribution of a continuous random variable can be represented by a probability density function (PDF).*
- *The expected value of a random variable is the weighted average of its possible values, with the weights given by their respective probabilities. It is a measure of the central tendency of the variable's probability distribution. The variance of a random variable measures how much its values deviate from its expected value, and it is a measure of the variability of the variable's probability distribution.*
- *In Python, we can use libraries like NumPy and SciPy to work with random variables and their properties.*

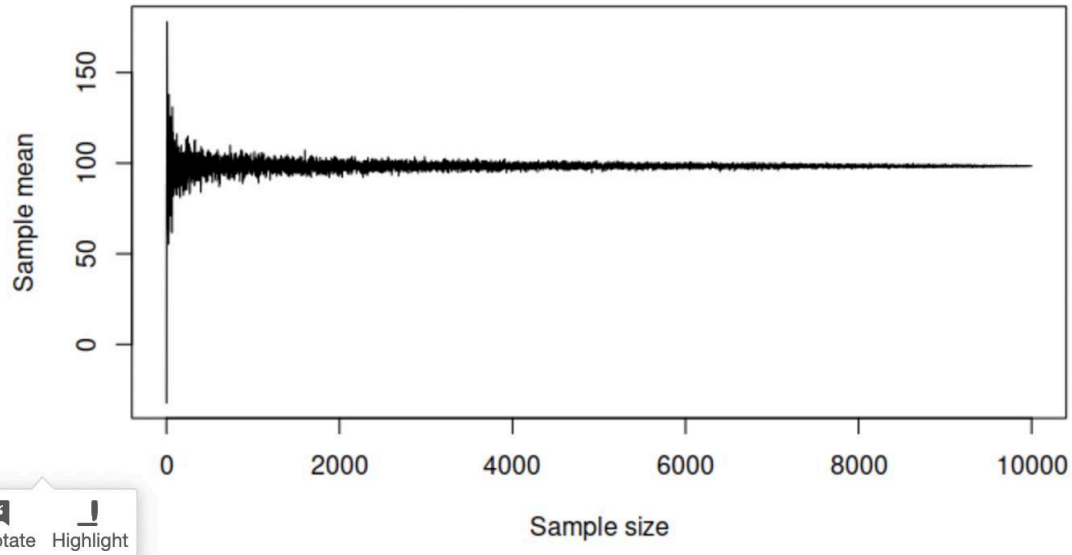
12. What is the law of large numbers, and how does it relate to probability theory?(20)

The ***law of large numbers*** *has a very central role in probability and statistics. The relative frequency interpretation of probability is that if an experiment is repeated a large number of times under identical conditions and independently, then the relative frequency with which an event A actually occurs and the probability of A should be approximately the same. A mathematical expression of this interpretation is the law of large numbers. This theorem says that if $X_1, X_2, \ldots, X_n$ are independent random variables having a common distribution with mean μ, then for any number $\varepsilon > 0$, no matter how small, as $n \to \infty$, the average of a variable*

obtained over the large number of trials will be close to its expected value and will get closer to it.

The law of large numbers is a fundamental concept in probability theory that describes the relationship between the sample size of a random variable and its expected value.

1. *The law of large numbers states that as the sample size of a random variable increases, the sample mean will approach the expected value of the variable. This means that the more data you have, the more accurate your estimate of the true underlying probability distribution will be.*
2. *This law applies to both discrete and continuous random variables, and it is a key concept in many areas of statistics and machine learning.*
3. *The law of large numbers is closely related to the central limit theorem, which states that the distribution of the sample means approaches a normal distribution as the sample size increases.*
4. *The law of large numbers has important implications for decision-making and risk management. It suggests that making decisions based on a large sample size is generally more reliable and accurate than making decisions based on a small sample size.*
5. *In practice, the law of large numbers is often used in simulations and statistical modeling to generate more accurate estimates of probabilities and other statistical measures. For example, if we want to estimate the probability of a rare event occurring, we can use the law of large numbers to simulate many trials and calculate the proportion of trials in which the event occurs.*



13. What is the central limit theorem, and how is it used?(21)

The Central Limit Theorem states that the sample mean follows a normal distribution as the sample size increases. In other words, when the sample size is large, the distribution of the sample mean will be normal regardless of the original distribution of the population, be it normal, Poisson, binomial, or any other type.

*The central limit theorem applies to almost all types of **probability distributions**, but there are exceptions. For example, the population must have a finite variance. That restriction rules out the Cauchy distribution because it has an infinite variance.*

Additionally, the central limit theorem applies to independent, identically distributed variables. In other words, the value of one observation does not depend on the value of another observation. And the distribution of that variable must remain constant across all measurements.

Conditions of the Central Limit Theorem:

- *The sample size is **sufficiently large**. This condition is usually met if the size of the sample is $n \geq 30$.*
- *The samples are **independent and identically distributed, i.e., random variables**. The sampling should be random.*
- *The population's distribution has a **finite variance**. The central limit theorem doesn't apply to distributions with infinite variance.*

The central limit theorem has many applications in different fields.

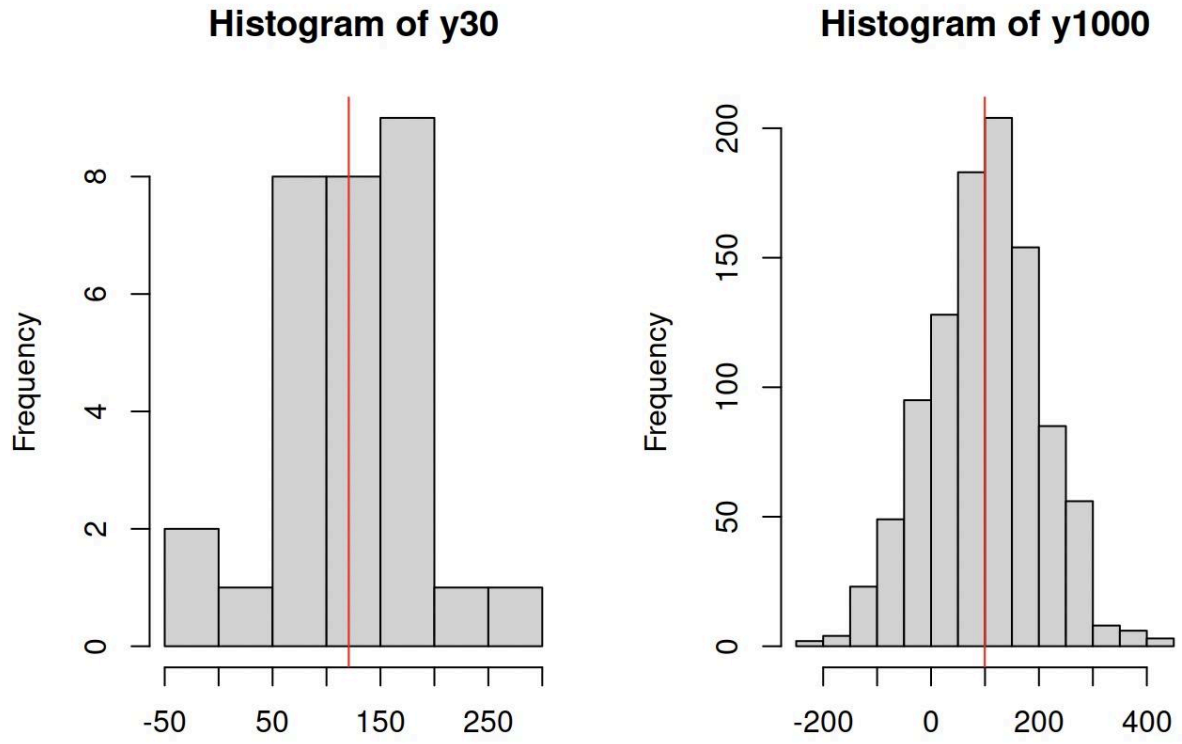
1. Political/election polls are prime CLT applications. These polls estimate the percentage of people who support a particular candidate. You might have seen these results on news channels that come with confidence intervals. The central limit theorem helps calculate the same.

2. Confidence interval, an application of CLT, is used to calculate the mean family income for a particular region.

Assumptions Behind the Central Limit Theorem

- *The **data must follow the randomization condition**. It must be sampled randomly*
- ***Samples should be independent of each other.** One sample should not influence the other samples*
- ***Sample size should be not more than 10% of the population** when sampling is done without replacement*
- *The **sample size should be sufficiently large**. Now, how will we figure out how large this size should be? Well, it depends on the population. When the population is skewed or asymmetric, the sample size should be large. If the population is symmetric, then we can draw small samples as well.*

In general, ***a sample size of 30 is considered sufficient when the population is symmetric****.*



14. What is the difference between discrete and continuous probability distributions? (22)

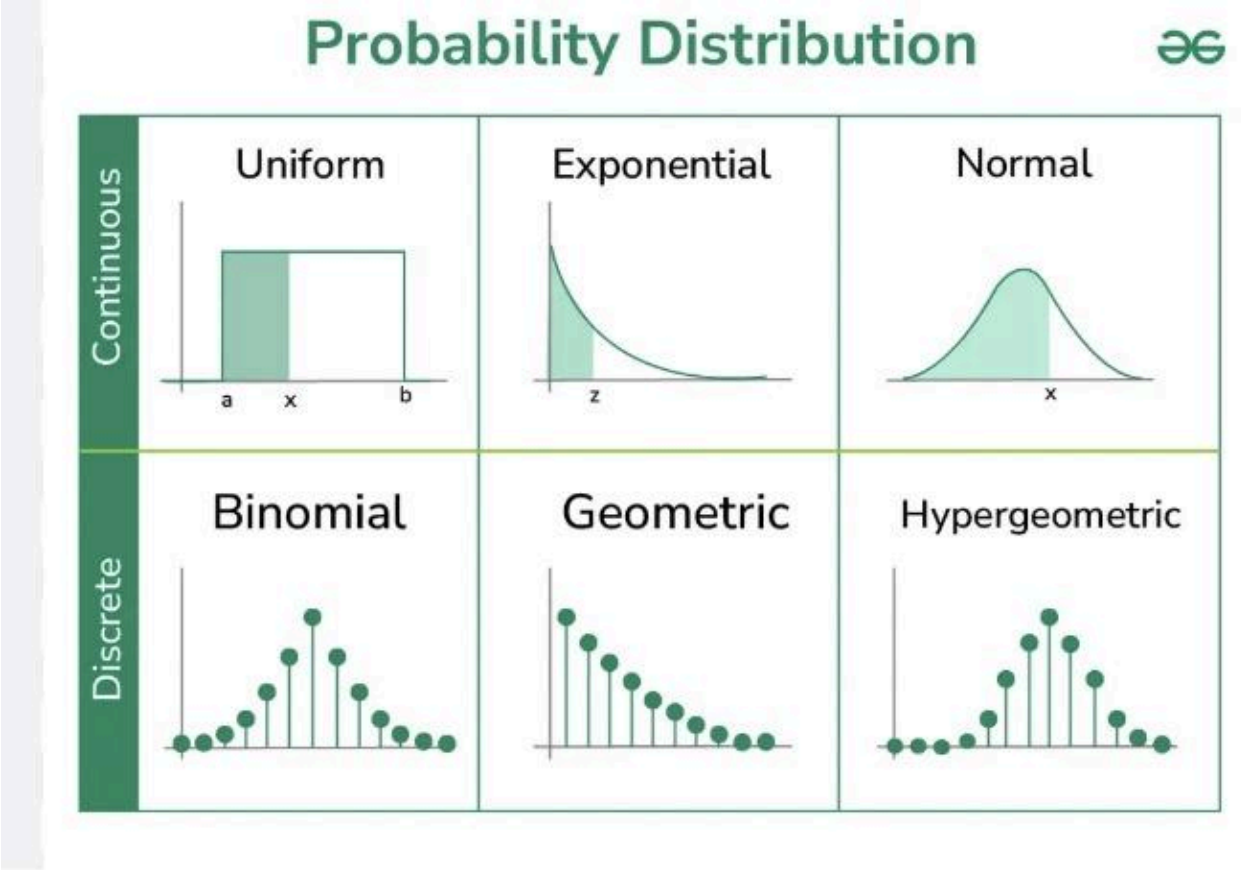
A ***discrete probability distribution*** *is a statistical function that describes the likelihood of obtaining a particular value or set of values from a discrete set of possible values distribution, such as the outcome of dice rolls, coin flips, the number of customers arriving at a service point, or the success/failure of a series of trials and are represented using the probability mass function, a function that maps each possible outcome of a discrete random variable to its probability of occurrence*

Examples of discrete probability distributions include the binomial distribution, the Poisson distribution, and the geometric distribution.

Unlike discrete distributions, which deal with countable outcomes, ***Continuous Probability Distributions*** *address variables that can take on any value within a given range, such as the behavior of particles, stock prices, and population growth. As a result, continuous probability distributions are represented by a probability density function (PDF).*

The area under the PDF curve between two points represents the probability of the random variable taking a value between those two points. The PDF is continuous and non-negative and its total area under the curve is equal to 1.

Some examples of continuous probability distributions are Normal Distribution, Uniform Distribution, Exponential Distribution, Beta Distribution, Gamma Distribution, and Weibull Distribution.



15. What are some common measures of central tendency, and how are they calculated?(23)

There are various measures of central tendency, such as mean, median, and mode, to describe the typical value of a dataset.

1. Mean:

It is the arithmetic average of a dataset and is calculated by summing all values in the dataset and dividing by the number of observations. Mean can be sensitive to outliers and extreme values in a dataset. The formula for calculating the mean is:

mean = (sum of all values) / (number of observations)

However, there are three types of means used in different scenarios, which include 1.Geometric Mean - The nth root of the product of all the values in a data set. It's used to compare growth rates or ratios. ***For example,*** *growth of population, deposits in a bank account attracting compound interest, depreciation charged using diminishing balance method, etc.*

2. Harmonic Mean - The reciprocal of the arithmetic mean of the reciprocals of the data set. It's used to calculate rates or reciprocals, and is especially useful when smaller values in the data set need more emphasis. It is used in the financial and insurance industries for risk management, etc.

3. Weighted Mean: A customized mean that takes into account the relative importance of each value in the data set. It's used when different data points have different importance.

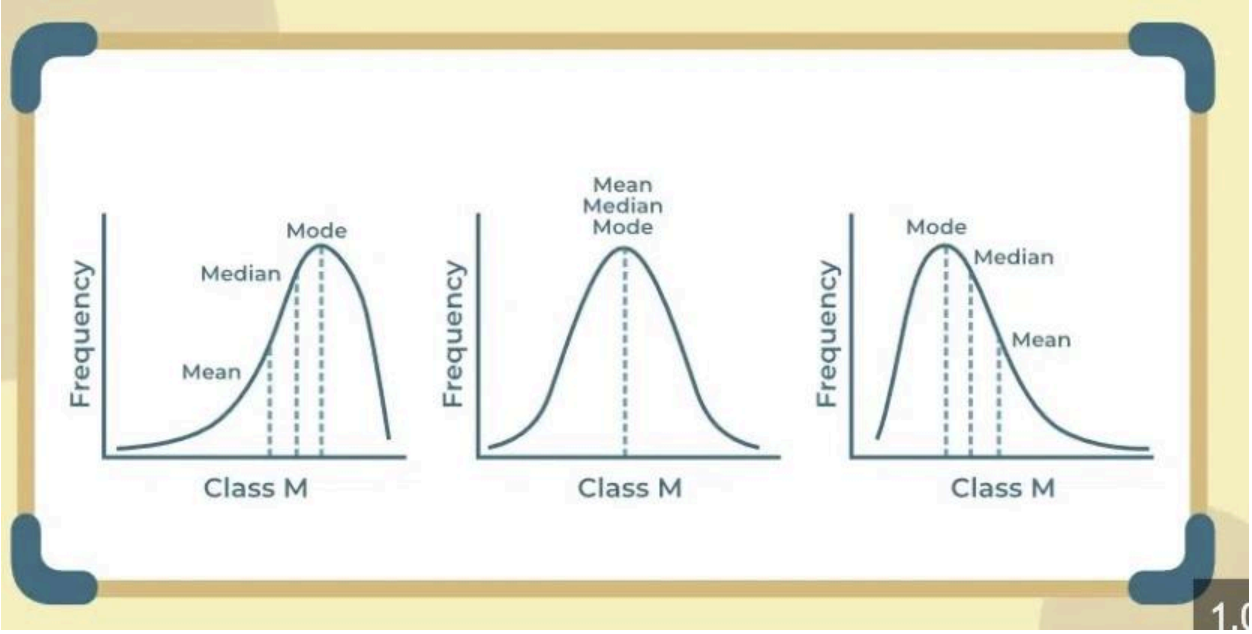
2*.* ***Median****:*

It is the middle value in a dataset when the values are arranged in ascending or descending order. Median is less sensitive to outliers compared to mean. In case of even number of observations, the median is calculated as the average of the two middle values.

3. Mode:

It is the most frequently occurring value in a dataset. Mode can be used for both numerical and

Mean, median, and mode are statistical measures that describe the main features of a dataset, including measures of central tendency.



16. What is the purpose of using percentiles and quartiles in data summarization

Percentiles and quartiles provide information about the distribution and and spread of data especially for large datasets.

Percentiles

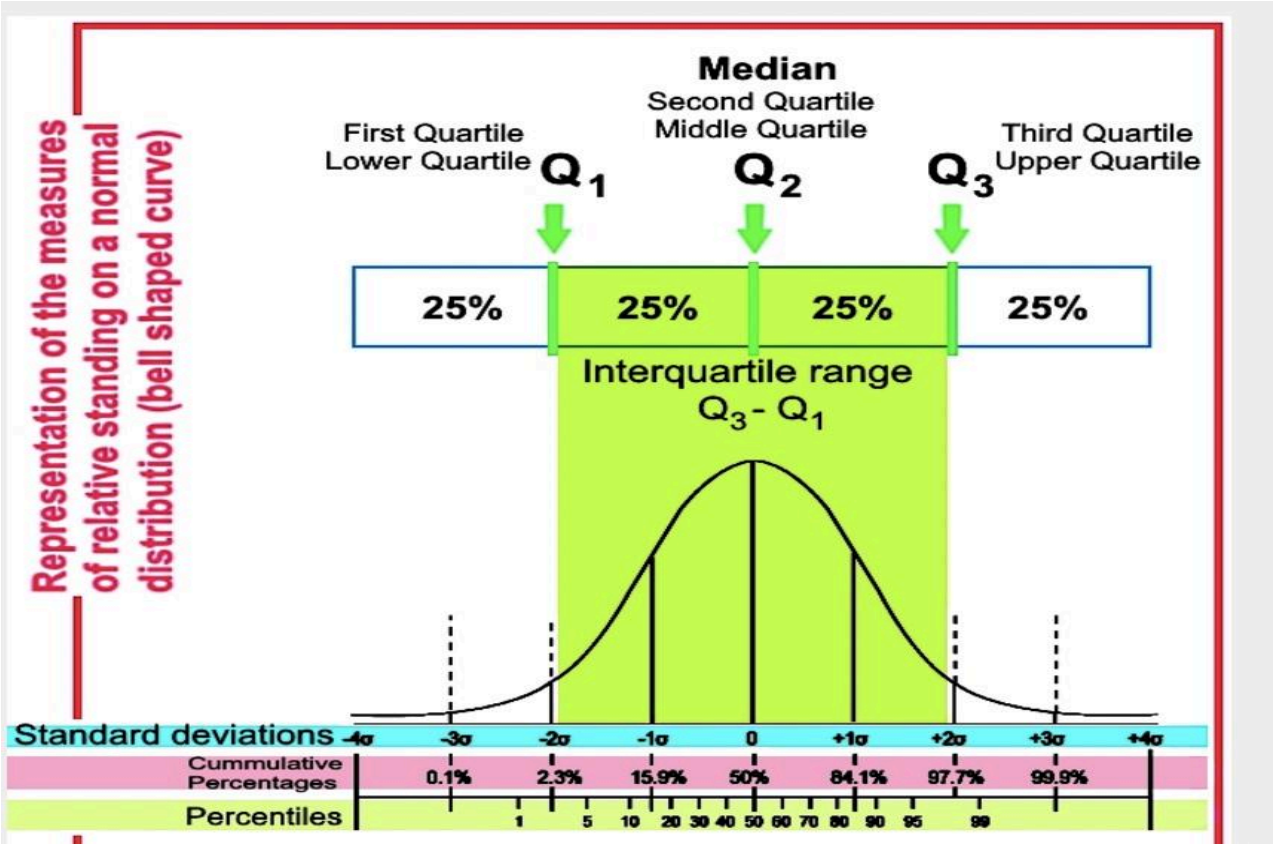
Help understand the rank of a data point compared to the rest of the data. For example, a data point at the 50th percentile is greater than 50% of the other data points. Percentiles divide a dataset into 100 equal parts, each representing 1% of the data. For example, the 75th percentile is the value below which 75% of the data falls.

Quartiles

Divide data into four equal parts, making it easier to understand the spread of data and identify outliers. The first quartile (Q1) is the 25th percentile, the second quartile (Q2) is the 50th percentile, and the third quartile (Q3) is the 75th percentile.

Percentiles and quartiles are calculated by sorting the data and determining the positions of the required points.

- *The most commonly used method for calculating percentiles and quartiles is the interpolation method. This method estimates the percentile value by interpolating between the two nearest values in the dataset.*
- *In Python, you can use the NumPy library to calculate percentiles and quartiles. The percentile() function can be used to calculate any percentile value, and the quantile() function*
  be used to calculate quartiles.



17. What is a joint probability distribution?(28)

Joint distribution is an important concept in probability theory and statistics, and it refers to the distribution of two or more random variables together. Joint probability mass function and joint cumulative distribution function are used to analyze the joint behavior of discrete random variables.

Joint probability mass function: *If we have two discrete random variables, their joint distribution can be described by a joint probability mass function, which gives the probability of each possible combination of values for the two variables.*

Joint probability density function: *If we have two continuous random variables, their joint distribution can be described by a joint probability density function, which gives the probability density at each point in the joint space.*

In order for joint probability to work, both events must be independent of one another, which means they aren't conditional or don't rely on each other
The joint probability distribution must satisfy certain properties, such as non-negativity and the sum of all probabilities equaling 1.

Understanding joint probability is crucial for various statistical analyses, including regression analysis, hypothesis testing, and decision-making.

18. What is the difference between a joint probability distribution and a marginal probability distribution?(30)

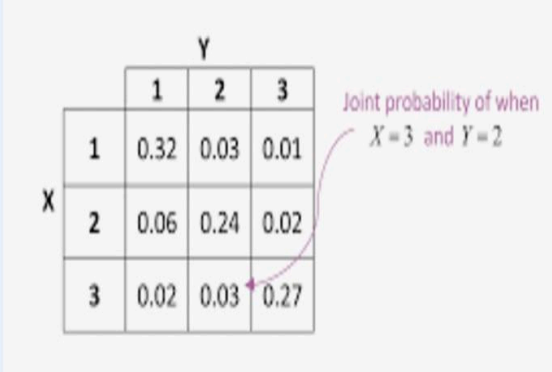
Joint distribution is an important concept in probability theory and statistics, and it refers to the distribution of two or more random variables together which is Joint probability mass function for discrete variables and Joint probability density function for continuous variables.
The marginal distribution is obtained by summing (or integrating, in the case of continuous variables) the joint probability distribution over the variables not of interest.Marginal distributions can be calculated for both discrete and continuous variables.

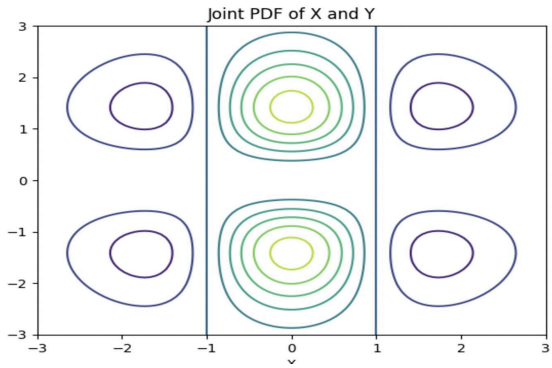
Joint distributions are used to model the relationships between multiple random variables and make predictions or decisions based on those relationships whereas Marginal distributions are important in statistics, as they allow us to study the behavior of individual variables in a multivariate distribution.
Marginal distributions are important in statistics, as they allow us to study the behavior of individual variables in a multivariate distribution.

The marginal distribution of a single variable can be obtained by summing (or integrating) the joint distribution over all possible values of the other variables.

The marginal distribution of multiple variables can be obtained by summing (or integrating) the joint distribution over all possible values of the variables not of interest.

In Python, the marginal distribution can be calculated using the numpy.sum() function for discrete variables and scipy.integrate.simps() function for continuous variables.



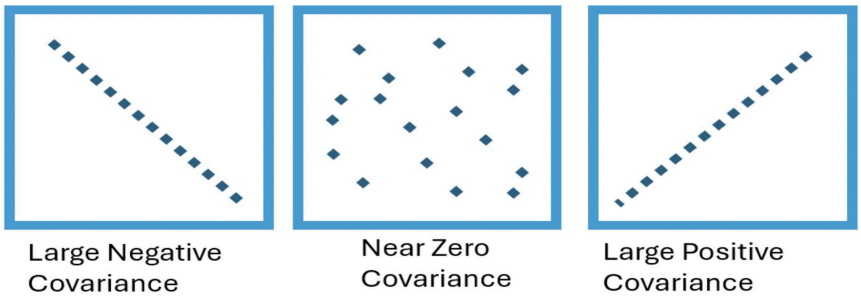


19. What is the covariance of a joint probability distribution?(31)

Covariance is a measure of how two variables change together. In the context of a joint probability distribution, it quantifies the linear relationship between these variables.Covariance can be calculated using the following formula:

$$Cov(X,Y)=E[(X-E[X])(Y-E[Y])]$$

A positive covariance means that the two variables tend to move in the same direction, while a negative covariance means they tend to move in opposite directions. Covariance is sensitive to the scale of the variables and is not standardized.



20. How do you determine if two random variables are independent based on their joint probability distribution?(32)

Two random variables, X and Y, are considered independent if and only if their joint probability distribution can be expressed as the product of their individual (marginal) probability distributions.

From a (more technical) standpoint, two random variables are independent if ***either of the following statements are true****:*

1. *P(x|y) = P(x), for all values of X and Y.*
2. *P(x∩y) = P(x) * P(y), for all values of X and Y.*

(OR)

P(X = x, Y = y) = P(X = x) P(Y = y), for all values of x and y.

The first statement, P(x|y) = P(x), for all values of X and Y, *is stating "the probability of x, given y, is x."*

You may recognize the second statement as the fundamental counting principle *, which states that if you have two independent events, multiply their probabilities together.*

21. What is the relationship between the correlation coefficient and the covariance of a joint probability distribution?(33)

Covariance is a measure of how two variables change together. In the context of a joint probability distribution, it quantifies the linear relationship between these variables.
Correlation measures the strength of the linear relationship between two variables. It ranges from -1 to 1, where -1 indicates a perfectly negative correlation, 0 indicates no correlation, and 1 indicates a perfectly positive correlation.
Covariance indicates the direction of the linear relationship between variables.
While Correlation measures both the strength and direction *of the linear relationship between two variables.*
Correlation Coefficient *between two random variables, X and Y*

$$Corr(X,Y) = Cov[X,Y] / (StdDev(X) . StdDev(Y)) .$$

Correlation is standardized, which means it is not sensitive to the scale of the variables.In other words, while covariance provides information about the linear relationship between variables, it's sensitive to the scale of the variables*. To obtain a scale-invariant measure of the relationship, use the correlation coefficient.*

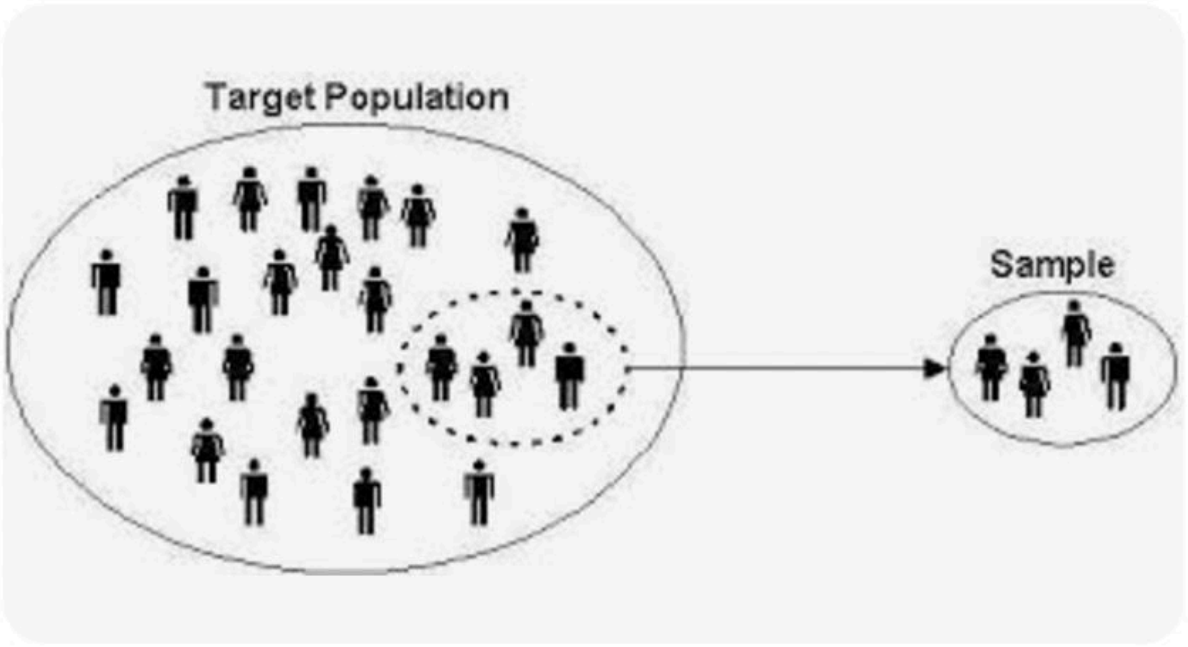
22. What is sampling in statistics, and why is it important?(34)

Because it's often not possible to measure the entire population, researchers use a representative sample to analyze. Sampling refers to the process of selecting a subset, known as a sample, from a larger group, known as a population. The sample is carefully chosen to be representative of the population's characteristics, allowing researchers to draw meaningful conclusions about the entire population based on the analysis of the sample data. Sampling involves techniques that aim to minimize bias and ensure the sample accurately reflects the diversity and variability present in the population.

Importance of Sampling in Statistics

1. ***Feasibility:*** *Studying an entire population is often impractical, expensive, or even impossible due to time, cost, and logistical constraints.*
2. ***Efficiency:*** *Sampling allows researchers to gather data more quickly and efficiently than studying the entire population.*
3. ***Cost-effectiveness:*** *Sampling reduces the costs associated with data collection, analysis, and processing*

4. ***Reduced Data Collection Effort:*** *Instead of gathering data from all individuals, researchers can focus on collecting data from a smaller group, making data collection more manageable.*
5. ***Risk Reduction:*** *Sampling allows researchers to evaluate hypotheses and test new ideas on a smaller scale before implementing them on the entire population, reducing potential risks and errors.*
6. ***Ethics:*** *In cases where it is not feasible or ethical to collect data from every individual, sampling provides a way to gather relevant information without invading privacy or causing harm.*



23. What are the different sampling methods commonly used in statistical inference?(35)

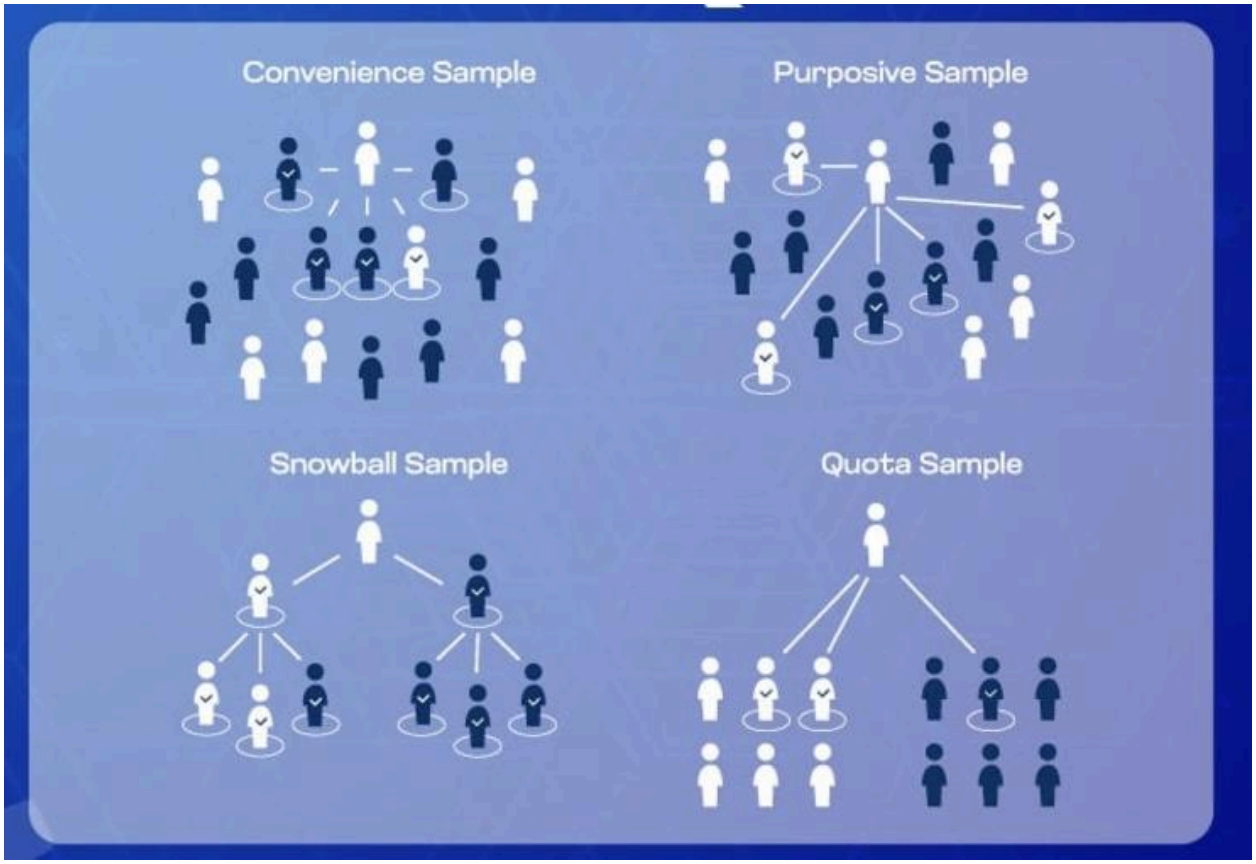
Probability Sampling (Random Sampling): *Every member of the population has a known and non-zero probability of being selected.*

- *Simple Random Sampling: Each member of the population has an equal chance of being selected. (e.g., drawing names from a hat)*

- *Stratified Random Sampling: The population is divided into subgroups (strata), and a random sample is taken from each stratum. (e.g., sampling students from each grade level in a school).*
- *Systematic Sampling: Selecting members at regular intervals from an ordered list of the population. (e.g., selecting every 10th person on a list)*
- *Cluster Sampling: Dividing the population into clusters and randomly selecting entire clusters for inclusion in the sample. (e.g., randomly selecting a few cities and surveying all residents within those cities)*

Non-Probability Sampling: *The selection of members is not based on random chance.*

- *Convenience Sampling: Selecting individuals who are readily available or easy to reach. (e.g., surveying customers in a store)*
- *Quota Sampling: Selecting a predetermined number of individuals from each subgroup within the population.*
- *Snowball Sampling: Identifying a few initial participants and then asking them to refer other potential participants. (often used in studies of hard-to-reach populations)*
- *Judgment or purposive or deliberate sampling: Judgment sampling, also known as purposive or deliberate sampling, involves selecting specific individuals or cases based on the researcher's judgment or specific criteria.*



24. What is the p-value in hypothesis testing?(38)

The P-value (or p-value or probability value) in hypothesis testing represents the probability of obtaining the observed results (or more extreme results) if the null hypothesis were true.

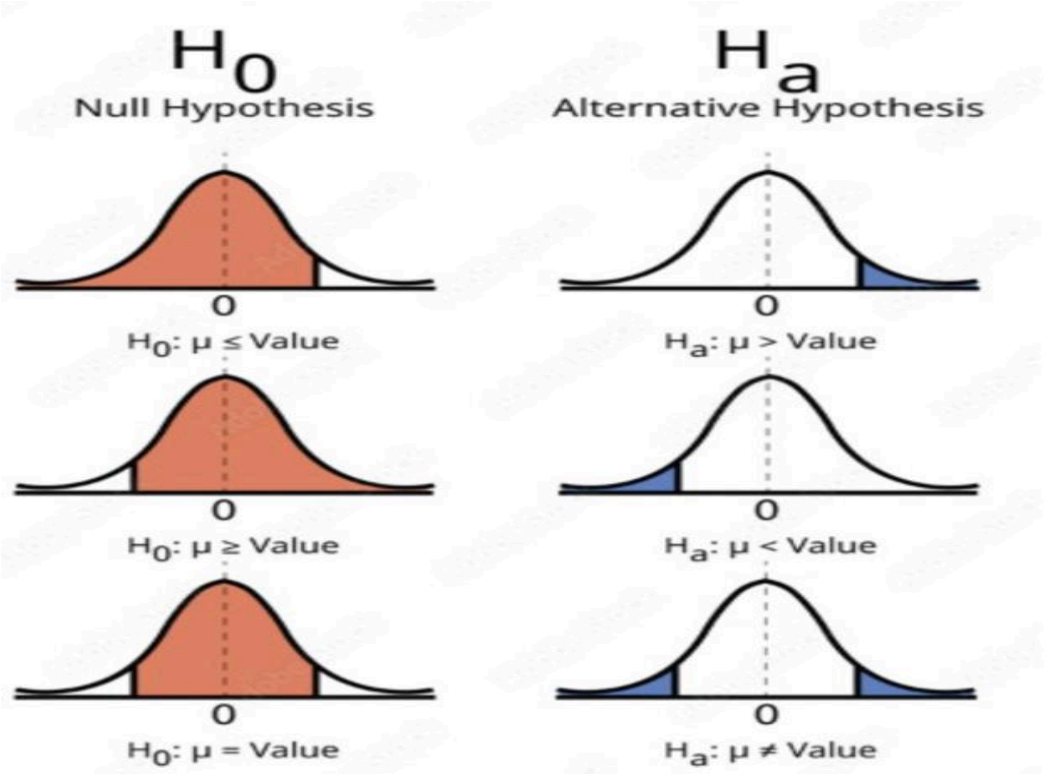
In simpler terms, imagine you have a coin and want to test if it's fair (i.e., has a 50/50 chance of landing on heads or tails).The null hypothesis would be: "The coin is fair."You flip the coin 100 times and get 80 heads. This seems unlikely if the coin is truly fair.The p-value tells you how likely it is to get 80 or more heads in 100 flips if the coin were actually fair.

We always test the null hypothesis. The p-value is not the probability that the null hypothesis is true.A common threshold for statistical significance is a p-value of 0.05, but this can vary depending on the field and the specific research question.

The initial conclusion will always be one of the following:

- *Reject the null hypothesis - if the P-value ≤ α (where α is the significance level, such as 0.05): This suggests that the observed results are unlikely to have occurred by chance if the null hypothesis were true. You would reject the null hypothesis.*

- *Fail to reject the null hypothesis - if the P-value > α -> This suggests that the observed results could have occurred by chance if the null hypothesis were true. You would fail to reject the null hypothesis.*



25. What are Type I and Type II errors in hypothesis testing?(40)

Type - I error: *A Type I error is the mistake of rejecting the null hypothesis when it is true.The symbol α (alpha) is used to represent the probability of a type I error*

Type - II error*: A Type II error is the mistake of failing to reject the null hypothesis when it is false.The symbol β (beta) is used to represent the probability of a type II error*

26. What is the difference between correlation and causation?(41)

In statistical analysis, correlation coefficients are a quantitative assessment that measures both the direction and the strength of this tendency to vary together. There are different types of correlation coefficients that you can use for different kinds of data.

A positive correlation example is the relationship between the speed of a wind turbine and the amount of energy it produces. As the turbine speed increases, electricity production also increases.

A negative correlation example is the relationship between outdoor temperature and heating costs. As the temperature increases, heating costs decrease.

However, Correlation Does Not Imply Causation. *Correlation between two variables indicates that changes in one variable are associated with changes in the other variable. However, correlation does not mean that the changes in one variable actually cause the changes in the other variable.*

Sometimes it is clear that there is a causal relationship. For the height and weight data, it makes sense that adding more vertical structure to a body causes the total mass to increase. Or, increasing the wattage of lightbulbs causes the light output to increase.

However, in other cases, a causal relationship is not possible. For example, ice cream sales and shark attacks have a positive correlation coefficient. Clearly, selling more ice cream does not

cause shark attacks (or vice versa). Instead, a third variable, outdoor temperatures, causes changes in the other two variables.

27. How is a confidence interval defined in statistics?(42)

A confidence interval (CI) is a range of values that is likely to contain the value of an unknown population parameter. These intervals represent a plausible domain for the parameter given the characteristics of your sample data.

***A confidence interval indicates where the population parameter is likely to reside**. For example, a 95% confidence interval of the mean [9 11] suggests you can be 95% confident that the population mean is between 9 and 11.*

The precise formula depends on the type of parameter you're evaluating. You'll use critical Z-values or t-values to calculate your confidence interval of the mean.

Where:

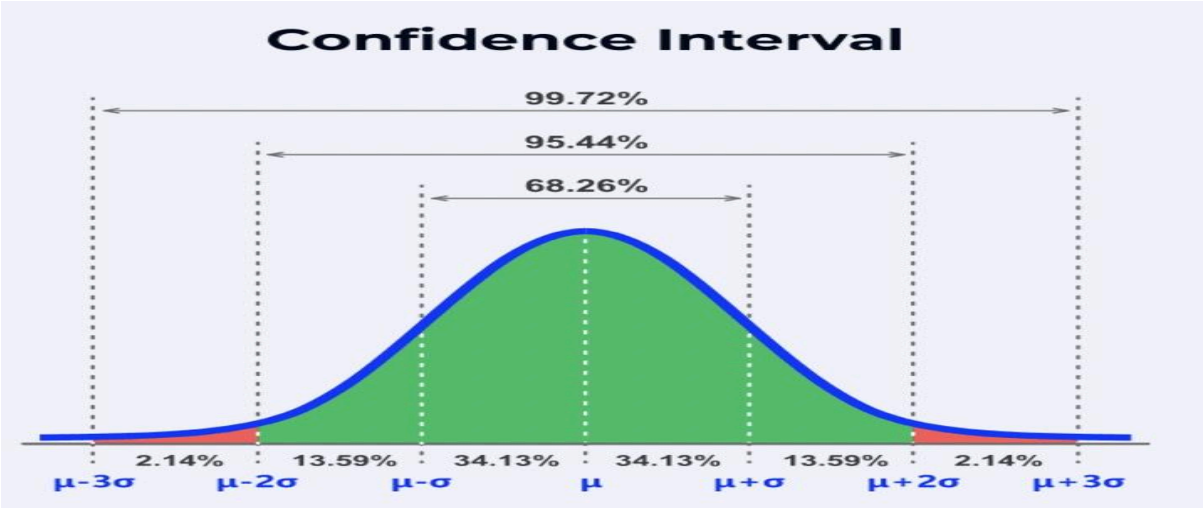
- $\bar{x}$ *= the sample mean, which is the point estimate.*
- *Z = the critical z-value*
- *t = the critical t-value*
- *s = the sample standard deviation*
- $s / \sqrt{n}$ *= the standard error of the mean*

To calculate a confidence interval, take the critical value (Z or t) and multiply it by the standard error of the mean (SEM). This value is known as the margin of error (MOE). Then add and subtract the MOE from the sample mean ($\bar{x}$) to produce the upper and lower limits of the range.

The width of a confidence interval is affected by sample size, confidence level, and standard deviation-

- *Increasing your sample size is the primary way to reduce the widths of confidence intervals because, in most cases, you can control it more than the variability.*
- *If you increase the confidence level (e.g., 95% to 99%) while holding the sample size and variability constant, the confidence interval widens.*
- *Variability present in your data affects the precision of the estimate. Your confidence intervals will be broader when your sample standard deviation is high.*

If a confidence interval includes the value of zero, it means that the difference is not statistically significant at the given confidence level. In hypothesis testing, the null hypothesis typically states that there is no difference between groups, and if the confidence interval contains zero, we fail to reject the null hypothesis.



28. What does the confidence level represent in a confidence interval?(43)

The confidence level is the probability that a sample statistic (e.g., a mean or proportion) falls within a given range of values. Imagine estimating the average height of people in your city. Confidence levels give you a range of possible values and tell you how sure you are that the true mean falls within that range.

The width of a confidence interval is affected by sample size, confidence level, and standard deviation.

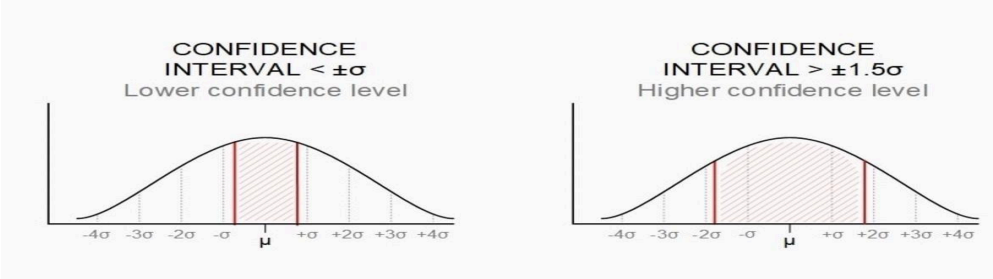
The confidence level also affects the confidence interval width. However, this factor is a methodology choice separate from your sample's characteristics.

If you increase the confidence level (e.g., 95% to 99%) while holding the sample size and variability constant, the confidence interval widens. Conversely, decreasing the confidence level (e.g., 95% to 90%) narrows the range.

Imagine you take your knowledge of a subject area and indicate you're 95% confident that the correct answer lies between 15 and 20. Then I ask you to give me your confidence for it falling between 17 and 18. The correct answer is less likely to fall within the narrower interval, so your confidence naturally decreases.

Conversely, I ask you about your confidence that it's between 10 and 30. That's a much wider range, and the correct value is more likely to be in it. Consequently, your confidence grows.

Confidence levels involve a tradeoff between confidence and the interval's spread. To have more confidence that the parameter falls within the interval, you must widen the interval. Conversely, your confidence necessarily decreases if you use a narrower range.



29. What is hypothesis testing in statistics?(44)

Hypothesis testing in statistics uses sample data to infer the properties of a whole population. These tests determine whether a random sample provides sufficient evidence to conclude an effect or relationship exists in the population.

For example:

- *StoreType A: 72 sales/day, StoreType B: 64 sales/day (n=10 days)*

→ Don't declare "Type A superior" without power analysis!

n=10 days? Pure gambling! 🎲 →small sample size results in Large standard error

- *Difference: 8 sales/day → t = 1.54, p = 0.12 (fail to reject H_0 ie Translation: "Insufficient evidence that StoreTypes differ")*

Days per Group	***Power (8-sales diff)***
10	*35% ❌*
30	*60% ⚠️*
85	***80% ✅***
176	***90% 🎯***

Verdict: Collect 3+ months data before declaring StoreType winners. Your full Rossmann dataset has this covered!-

calculated using TTestIndPower available in statsmodels.stats.power

Researchers use them to help separate genuine population-level effects from false effects that random chance can create in samples through significance testing.

A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement the sample data best supports. These two statements are called the null hypothesis and the alternative hypothesis.
We can use statistical tests to evaluate the likelihood that our hypothesis is true, or to identify other possible explanations for our observations

1. *Formulate the Hypotheses: Write your research hypotheses as a null hypothesis (H0) and an alternative hypothesis (HA).*
2. *Establish the power and significance level to calculate the sample size required*
3. *Data Collection: Gather data specifically aimed at testing the hypothesis.*
4. *Conduct A Test: Use a suitable statistical test to analyze your data and determine the p-value*
5. *Make a Decision: Based on the statistical test results, decide whether to reject the null hypothesis or fail to reject it.(If the p-value is less than or equal to alpha, you can reject the null hypothesis, indicating statistical significance.)*
6. *Report the Results: Summarize and present the outcomes in your report's results and discussion sections.*

The three major types of hypotheses are:

1. *Null Hypothesis (H0): Represents the default assumption, stating that there is no significant effect or relationship in the data.*
2. *Alternative Hypothesis (Ha): Contradicts the null hypothesis and proposes a specific effect or relationship that researchers want to investigate.*
3. *Nondirectional Hypothesis: An alternative hypothesis that doesn't specify the direction of the effect, leaving it open for both positive and negative possibilities.*

Selecting the right hypothesis test depends on several factors: the objective of your analysis, the type of data (numerical or categorical), and the sample size. Consider whether you're comparing means, proportions, or associations, and whether your data follows a normal distribution. The correct choice ensures accurate results tailored to your research question.

***Limitations**: Hypothesis testing has some limitations that researchers should be aware of:*

1. *It cannot prove or establish the truth: Hypothesis testing provides evidence to support or reject a hypothesis, but it cannot confirm the absolute truth of the research question.*
2. *Results are sample-specific: Hypothesis testing is based on analyzing a sample from a population, and the conclusions drawn are specific to that particular sample.*

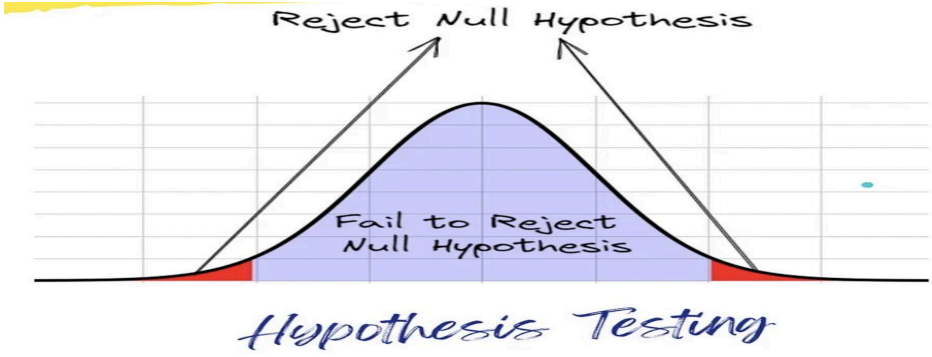
3. *Possible errors: During hypothesis testing, there is a chance of committing type I error (rejecting a true null hypothesis) or type II error (failing to reject a false null hypothesis).*
4. *Assumptions and requirements: Different tests have specific assumptions and requirements that must be met to accurately interpret results.*

30. What is the purpose of a null hypothesis in hypothesis testing?(45)

The null hypothesis (denoted by H_0) is a statement that the value of a population parameter (such as proportion, mean, or standard deviation) is equal to some claimed value. ***It is one of two mutually exclusive hypotheses about a <u>population</u> in a hypothesis test.***

- *We test the null hypothesis directly.*
- *Either reject H_0 or fail to reject H_0.*
- *The null hypothesis is what we are willing to assume is the case until proven otherwise. We can never claim that the null hypothesis has been actually proved.*

The significance level (denoted by α) defines how much evidence we require to reject Ho in favor of Ha



31. What is the difference between a one-tailed and a two-tailed test?(46)

In hypothesis testing, we aim to determine if there's enough evidence from a sample to reject a claim about a population parameter. The main difference between a one-tailed and a two-tailed test is ***the number of critical regions and the direction of the test***

One-Tailed Test

- ***Focus:*** *Examines whether the population parameter is significantly* ***greater than*** *or* ***less than*** *the hypothesized value.*
- ***Directionality:*** *Specifies a direction of the effect (e.g., "greater than," "less than").*
- ***Critical Region:*** *The rejection region lies in only* ***one tail*** *of the distribution (either the left or right tail).*

Two-Tailed Test

- ***Focus:*** *Examines whether the population parameter is significantly* ***different*** *from the hypothesized value, without specifying a direction.*
- ***Directionality:*** *Does not specify a direction of the effect.*
- ***Critical Region:*** *The rejection region is divided into* ***two tails*** *of the distribution (both left and right).*

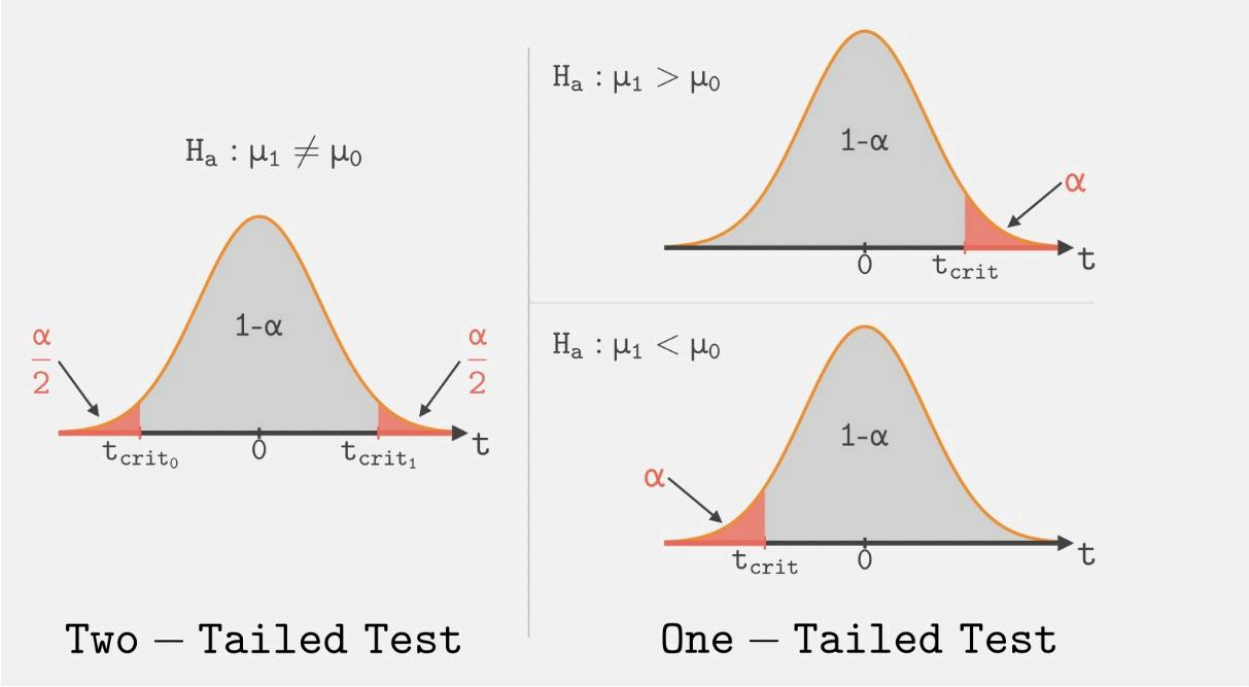
When to Use Which Test

- ***One-tailed test:*** *Use when you have a strong prior expectation about the direction of the effect.*
- ***Two-tailed test:*** *Use when you don't have a strong prior expectation about the direction of the effect or when you're interested in detecting any significant difference.*

Example

- ***One-tailed test:*** *The manufacturer now decides that it is only interested whether the mean lifetime of an energy-saving light bulb is less than 60* ***Two-tailed test:*** *A light bulb manufacturer claims that its' energy saving light bulbs last an average of 60 days.*

In summary: *One-tailed tests are more specific about the expected direction of the effect, while two-tailed tests are more general. The choice between a one-tailed and two-tailed test depends on the research question and the prior knowledge about the phenomenon being studied.*



32. What is the geometric interpretation of the dot product?(51)

The dot product of two vectors has a powerful geometric interpretation:

Projection and Magnitude: *Imagine projecting one vector onto the other.[1] The dot product is equal to the magnitude of one vector multiplied by the magnitude of the projection of the other vector onto it.*

Formula: $a \cdot b = |a| * |b| * \cos(\theta)$

where:

a and b are the vectors[3]

|a| and |b| are their magnitudes

θ is the angle between the vectors

Angle Between Vectors: *The dot product can be used to determine the angle between two vectors.[6]*

Formula: $\cos(\theta) = (a \cdot b) / (|a| * |b|)$

Example: *In physics, the dot product represents the work done by a force acting on an object.*

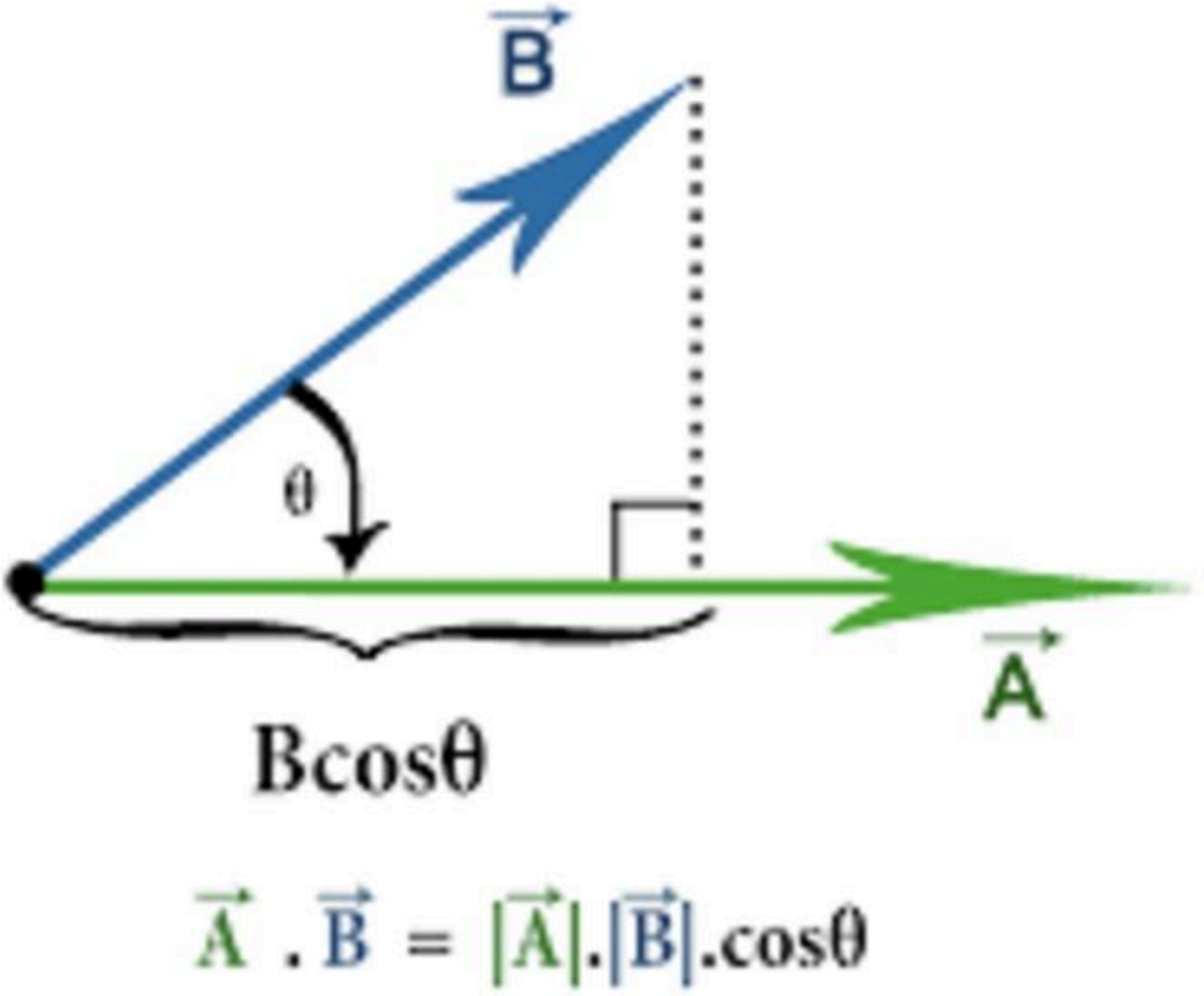
Work = Force · Displacement

The work is maximized when the force and displacement vectors are parallel (θ = 0°).

Key Points:

- *The dot product is a scalar quantity (a single number), not a vector.*
- *If the vectors are perpendicular (θ = 90°), the dot product is zero.*
- *If the vectors are parallel (θ = 0°), the dot product is the product of their magnitudes.*

By understanding these geometric interpretations, you can gain a deeper insight into the meaning and applications of the dot product in various fields, including physics, computer graphics, and machine learning.



33. What is the geometric interpretation of the cross-product?(52)

The cross product of two vectors, a and b, has a powerful geometric interpretation:
Perpendicular Vector and Area:

- *Perpendicularity: The cross product, a x b, results in a new vector that is perpendicular to both a and b.[1] This means it's orthogonal to the plane formed by the two original vectors.[2]*
- *Area: The magnitude of the cross product, |a x b|, is equal to the area of the parallelogram formed by the two vectors a and b as its adjacent sides.[3]*

Right-Hand Rule:

- *Direction: The direction of the resulting vector (a x b) is determined by the right-hand rule.[4] If you curl the fingers of your right hand from a towards b, your extended thumb will point in the direction of the cross product.[5]*

Key Points:

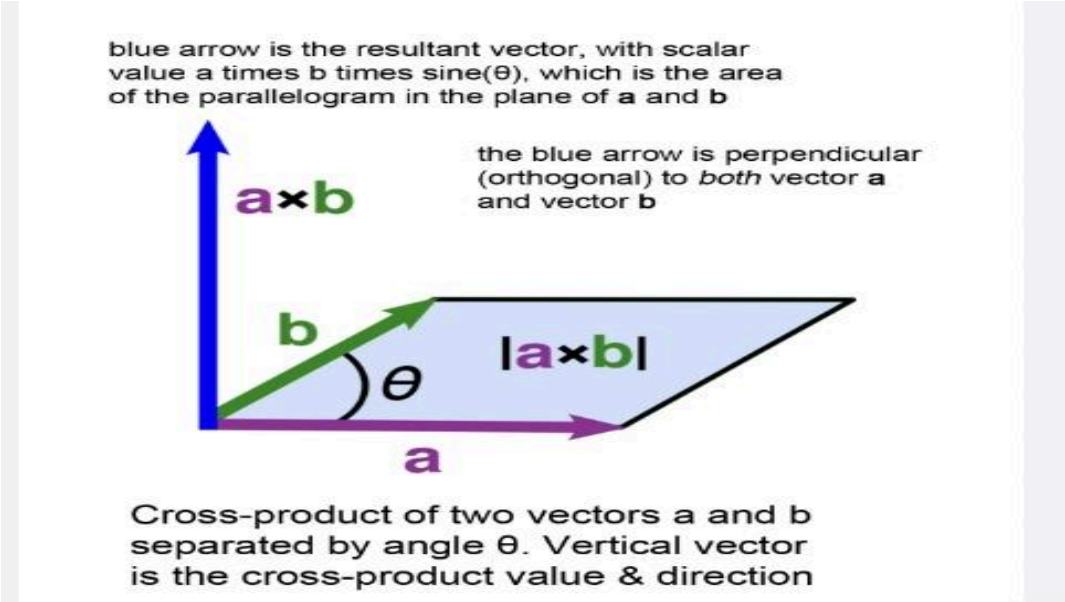
1. *The cross product is a vector quantity, unlike the dot product which is a scalar.[6]*
2. *If the two vectors are parallel or antiparallel (point in opposite directions), their cross product is zero.[7]*
3. *The cross product is anticommutative, meaning a x b = -b x a.[8]*

Examples*:*

In Physics, the torque of a force is calculated as the cross product of the force vector and the lever arm vector.

In data science, particularly in areas involving 3D data and geometric representations.

By understanding these geometric interpretations, you can gain a deeper insight into the meaning and applications of the cross product in various fields of physics, mathematics, and computer science.



34. How are optimization algorithms with calculus used in training deep learning models?

Deep learning models are trained by adjusting their parameters (weights and biases) to minimize a loss function. This optimization process often involves calculus-based algorithms. Here's how they are used:

- *Limits & Continuity (Foundation of Derivatives)*

  In ML and DL, functions must be continuous and smooth to allow efficient optimization.

  Imagine training a neural network with a ***discontinuous*** *loss function. When updating weights, the gradient might suddenly jump, making learning unstable.* ***Continuous functions ensure stable convergence.***

  Optimization algorithms with calculus are used in training deep learning models to ***minimize a loss function by iteratively updating model parameters using derivatives****.*

  During training, the model's error is measured by a ***loss function - these*** *functions must be continuous and smooth to allow efficient optimization.*

  Using ***calculus (partial derivatives)****, the optimizer computes the* ***gradient*** *of the loss with respect to each weight. This gradient indicates the direction of steepest increase in error, so the weights are updated in the* ***opposite direction*** *to reduce the loss.*

  Because deep neural networks are composed of many nested functions, the ***chain rule of calculus*** *is applied through* ***backpropagation*** *to efficiently compute gradients for all layers, from the output layer back to the input layer. So the* ***Essential Calculus Concepts***

- ***Derivatives****: A derivative measures the instantaneous rate of change of a function with respect to a single input. In deep learning, derivatives help determine how sensitive the loss function is to changes in a single model parameter, guiding the adjustment of that parameter to reduce the error.*
- ***Partial Derivatives****: In models with many parameters (common in deep learning), a partial derivative measures the effect of changing one parameter while holding all others constant.*
- ***Gradient****: The gradient is a vector that compiles all the partial derivatives of a multivariable function (the loss function) with respect to each of its parameters. It points in the direction of the steepest increase of the function.*

- ***Gradient Descent****: This is the fundamental optimization algorithm that uses the gradient. To minimize the loss, the algorithm takes small, iterative steps in the opposite direction of the gradient, moving "downhill" toward the lowest point (the minimum loss).*
- ***Chain Rule****: This is a vital rule for calculating derivatives of composite functions (functions within functions). In neural networks, which are essentially nested functions, the chain rule allows the efficient computation of gradients across all layers, from the output back to the input. This process is known as* ***backpropagation****.*
- ***Maxima and Minima****: Calculus provides the tools to find critical points (where the derivative is zero). In optimization, the goal is to find the global minimum of the loss function. Deep learning models often navigate complex, non-convex loss surfaces that include local minima and saddle points, which advanced optimization techniques help to address.*
- ***Hessian Matrix****: This matrix of second-order partial derivatives captures the curvature of the loss function. While computationally expensive for deep networks, it is used in more advanced, second-order optimization methods to converge faster in certain scenarios.*

Derivatives of a function at any point describe the slope of that function at that point.

- *A positive slope represents a condition where an increase in the value of x will increase the value of f(x). If we decrease the value of x, the value of f(x) will also decrease.*
- *A negative slope represents a condition where an increase in the value of x will decrease the value of f(x). If we decrease the value of x, the value of f(x) will increase.*

So, if f(x) is the cost function, x is the parameter, and the goal is to achieve that value of x, for which f(x) is the minimum.

Optimization algorithms start with a random value for x and then use the cost function's derivatives to understand whether it should increase or decrease the parameter's value to achieve the minimum.

In starting, the machine will randomly pick any 'θ1', and only three possible scenarios can be expected:

θ1 > θ1': The optimization algorithm will ask ML algorithms to reduce the value for θ1 so that θ1 → θ1'.

θ1 < θ1': The optimization algorithm will ask ML algorithms to increase the value for θ1 so that θ1 → θ1'.

θ1 = θ1': No update in the parameter value is required. ML process will be completed.

The update parameter uses a negative (-) of the derivative. Hence, in the case of a positive slope, the value for that parameter will decrease; if the slope is negative, the value for the parameter will increase.

***Note**: While calculating the partial derivative of a function with respect to any one parameter out of multiple parameters, we treat that parameter as the only variable affecting that function at that time. Rest all parameters are considered to be constant for that function. This helps to focus on one parameter during the updation process, and that's why we use partial derivatives in machine learning.*

1. Stochastic Gradient Descent (SGD):
*This is the fundamental building block. It relies on the **First Derivative** (the gradient) to find the slope of the loss function.*

The Calculus Function:

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla J(\theta_t)$$

- $\nabla J(\theta_t)$: This is the gradient (vector of partial derivatives). It tells the model which direction is "uphill." +1
- η **(Eta)**: The learning rate. Since the gradient points uphill, we subtract it to move downhill. +1

2. SGD with Momentum

*Standard SGD can get stuck in "local minima" (small dips) or oscillate in narrow valleys. Momentum uses calculus to add "velocity" to the movement, effectively calculating a **moving average** of the gradients.*

The Calculus Function:

1. $v_t = \gamma v_{t-1} + \eta \nabla J(\theta_t)$
2. $\theta_{t+1} = \theta_t - v_t$

- v_t: The velocity. By adding a fraction (γ) of the previous direction, the model gains "inertia," allowing it to push through small bumps in the loss landscape. +1

3. AdaGrad (Adaptive Gradient)

*In deep learning, some weights need to change more than others. AdaGrad uses calculus to track the **sum of squares of the gradients** to scale the learning rate for each parameter individually.*

The Calculus Function:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \epsilon}} \cdot \nabla J(\theta_t)$$

Gt*: The diagonal matrix where each element is the sum of the squares of the gradients up to time t.*

Logic*: If a weight has a very steep gradient (frequently updated), the denominator becomes large, effectively shrinking the learning rate to prevent overshooting.*

3. Advanced Optimizers:

- ***RMSprop:*** *Similar to AdaGrad but with a decaying average of past gradients to prevent the learning rate from becoming too small.*
- **Adam (Adaptive Moment Estimation):** *Adam is currently the "gold standard" in deep learning. It combines the ideas of* ***Momentum*** *(first moment) and* ***RMSProp*** *(second moment/variance). It uses calculus to estimate the mean and the uncentered variance of the gradients.*

The Calculus Functions:

1. **Mean (m_t):** $\beta_1 m_{t-1} + (1 - \beta_1)\nabla J(\theta_t)$
2. **Variance (v_t):** $\beta_2 v_{t-1} + (1 - \beta_2)(\nabla J(\theta_t))^2$
3. **Update:** $\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} \hat{m}_t$

Adam essentially calculates how much the gradient is "vibrating." If the gradient is consistent, it takes bigger steps. If the gradient is noisy/vibrating, it takes smaller, more cautious steps.

Different optimization algorithms extend this basic calculus idea:

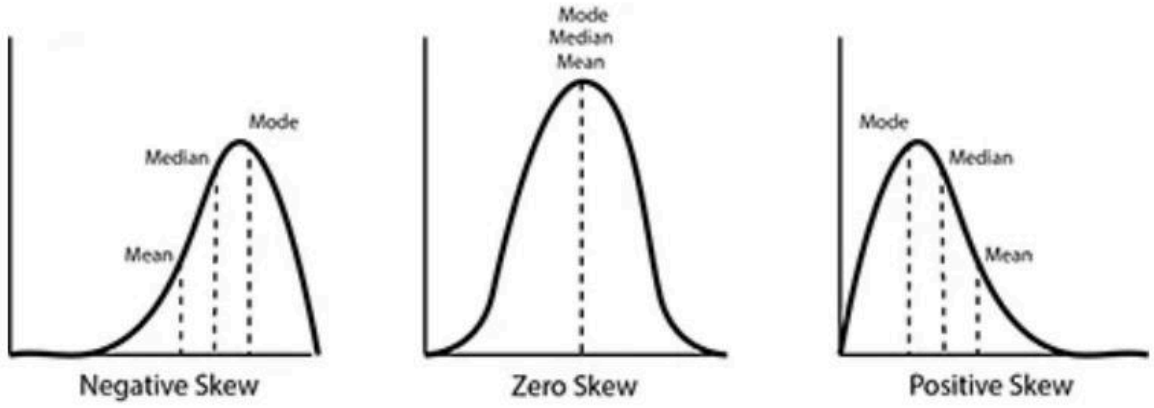
- ***SGD*** *uses first derivatives to take small steps downhill.*
- ***Momentum*** *accumulates past gradients to speed up convergence and reduce oscillations.*
- ***AdaGrad / RMSProp*** *use squared gradients to adapt learning rates.*
- ***Adam*** *combines momentum (first moment) and adaptive scaling (second moment) for faster and more stable convergence.*
-

In summary:
Calculus enables deep learning models to know how much and in which direction to change each weight, allowing optimization algorithms to efficiently train neural networks by minimizing prediction error.

35. What is the left-skewed distribution and the right-skewed distribution?(57)
Skewness refers to the degree of asymmetry in a distribution. Skewness can be calculated using the skew() function in the scipy.stats module in Python. The function takes an array of numbers as input and returns the skewness value.
***Right-Skewed Distribution (Positively Skewed)** : A distribution is said to be positively skewed if its tail is longer on the positive side (to the right) of the distribution. This means that the majority of the data is on the left side of the distribution.*

***Left-Skewed Distribution (Negatively Skewed):** A distribution is said to be negatively skewed if its tail is longer on the negative side (to the left) of the distribution. This means that the majority of the data is on the right side of the distribution. Mean < Median: The mean is typically less than the median. The mean is typically greater than the median.*

***Zero skewness:** A distribution is said to have zero skewness if it is symmetric around its mean. This means that the left and right tails are of equal length.*



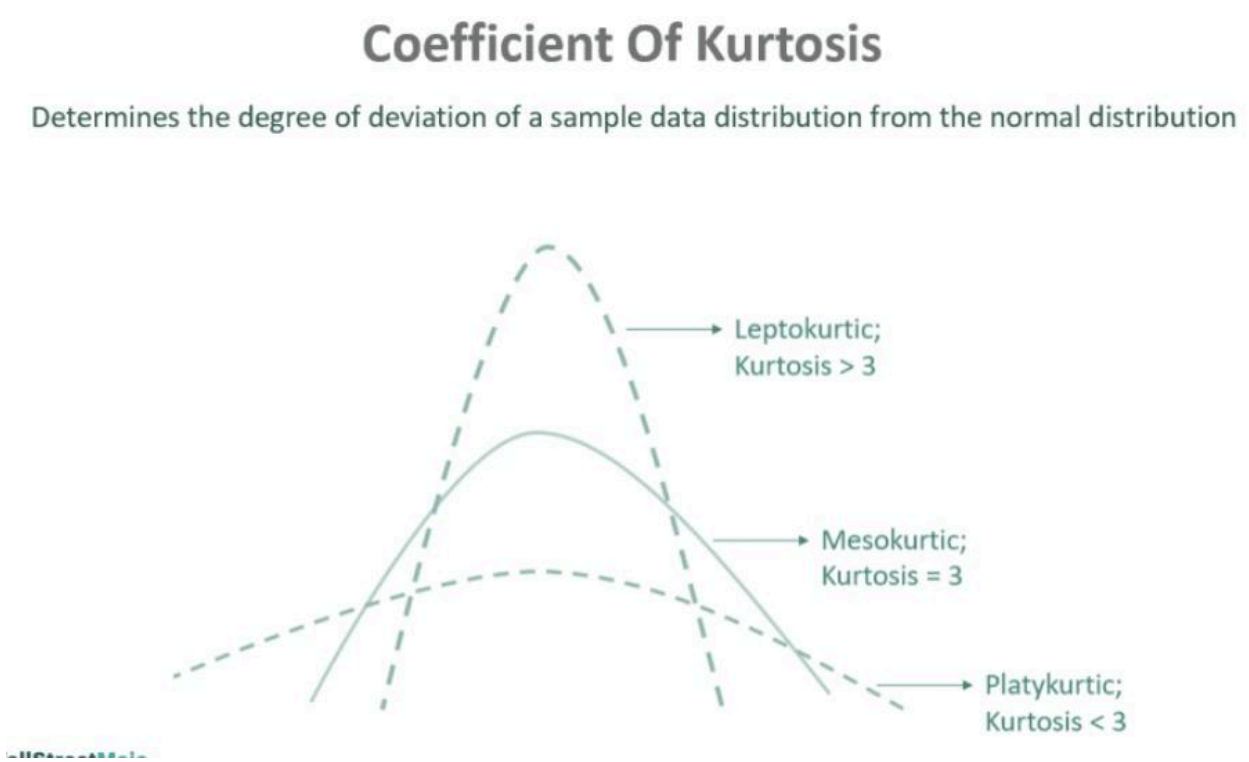
36. What is kurtosis?(58)
Kurtosis is a measures the degree of peakedness of a distribution and a statistic that measures the extent to which a distribution contains outliers. It assesses the propensity of a distribution to

have extreme values within its tails. There are three kinds of kurtosis: leptokurtic, platykurtic, and mesokurtic.

A high kurtosis indicates that the distribution has a sharper peak, heavier tails and hence more outliers falling relatively far from the mean, while a low kurtosis indicates a flatter peak, lighter tails and a lower tendency for producing extreme values.

When you're assessing a <u>sample</u>, outliers have the greatest impact on this statistic.

Excess kurtosis is a form of the statistic that helps you compare the tails of your distribution to those of the normal distribution. The excess form simply takes the standard statistic and normalizes it by subtracting 3.



37. What is the difference between Descriptive and Inferential Statistics?(60)

<u>Descriptive Statistics</u>

Descriptive statistics refers to a set of methods used to summarize and describe the main features of a dataset, such as its central tendency, variability, and distribution. These methods provide an overview of the data and help identify patterns and relationships. 1) providing basic information

about variables in a dataset and 2) highlighting potential relationships between. Various tools used are

***Central tendency**: Use the mean or the median to locate the center of the dataset. This measure tells you where most values fall.*

***Dispersion:** How far out from the center do the data extend? You can use the range or standard deviation to measure the dispersion. A low dispersion indicates that the values cluster more tightly around the center. Higher dispersion signifies that data points fall further away from the center. We can also graph the frequency distribution.*

***Skewness:** The measure tells you whether the distribution of values is symmetric or skewed. See: Skewed Distributions*

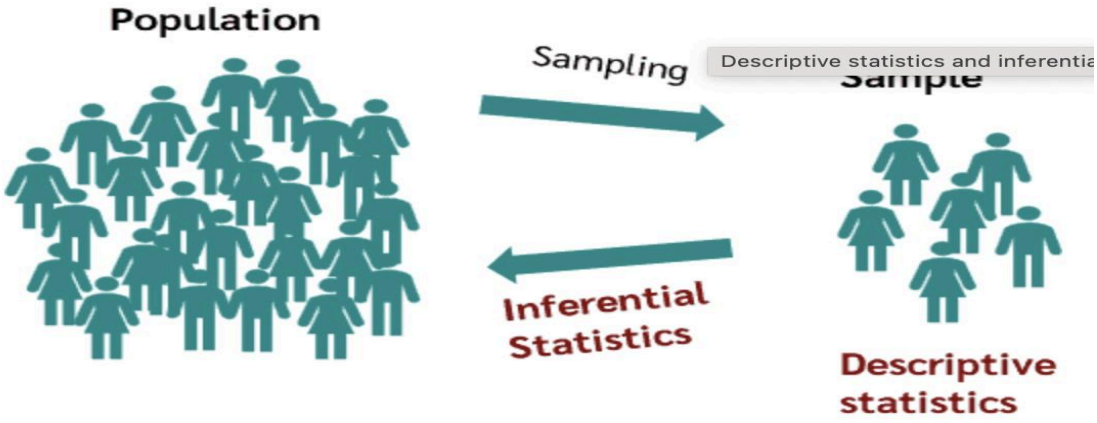
*These are the standard descriptive statistics, but there are other descriptive analyses you can perform, such as assessing the relationships of paired data using **correlation and scatterplots.***

Inferential Statistics

Inferential statistics takes data from a sample and makes inferences about the larger population from which the sample was drawn. Because the goal of inferential statistical analysis is to draw conclusions from a sample and generalize them to a population, we need to have confidence that our sample accurately reflects the population.

*The most common methodologies in inferential statistics are **hypothesis tests, confidence intervals, and regression analysis.** Interestingly, these inferential methods can produce similar summary values as descriptive statistics, such as the mean and standard deviation. However, we use them very differently when making inferences. Because it's often not possible to measure the entire population, researchers use a representative sample to analyze. However, sample statistics are unlikely to be exactly equal to the population value. Inferential statistics incorporate estimates of this error into the statistical results.*

A study using descriptive statistics is simpler to perform. However, if you need evidence that an effect or relationship between variables exists in an entire population rather than only your sample, you need to use inferential statistics.



38. What is the meaning of degrees of freedom (DF) in statistics?(63)

Degrees of freedom represent the number of independent values that can vary in a data sample*.Essentially, it's about how many values are free to change without violating any constraints or assumptions. It is an essential idea that appears in many contexts throughout statistics including* ***hypothesis tests, probability distributions, and linear regression.***

Degrees of freedom also define the probability distributions for the test statistics of various hypothesis tests. For example, hypothesis tests use the t-distribution, F-distribution, and the chi-square distribution to determine statistical significance. Each of these probability distributions is a family of distributions where the DF define the shape. Hypothesis tests use these distributions to calculate p-values. So, the DF directly link to p-values through these distributions!

In linear regression, the error DF are the independent pieces of information that are available for estimating your coefficients. For precise coefficient estimates and powerful hypothesis tests in regression, you must have many error degrees of freedom, which equates to having many observations for each model term

Calculating the degrees of freedom is often the sample size minus the number of parameters you're estimating:

$$DF = N - P$$

Where:

- ***N = sample size***
- ***P = the number of parameters or relationships***

Common Examples of Calculating Degrees of Freedom:

- ***One sample t-test:*** *DF = n - 1*
- ***Two sample t-test:*** *DF = n1 + n2 - 2*
- ***Simple linear regression:*** *DF = n - 2*
- ***Chi square goodness of fit test:*** *DF = k - 1*
- ***Chi square test for homogeneity:*** *DF = (r - 1)(c - 1)*

39. What is the empirical rule in Statistics?(65)

*The **Empirical Rule**, also known as the **68-95-99.7 Rule**, is a statistical guideline that describes the distribution of data in a normal distribution.*

Here's the breakdown:

- ***68%:*** *Approximately 68% of the data falls within one standard deviation (σ) of the mean (μ).*
- ***95%:*** *Approximately 95% of the data falls within two standard deviations (2σ) of the mean.*
- ***99.7%:*** *Approximately 99.7% of the data falls within three standard deviations (3σ) of the mean.*

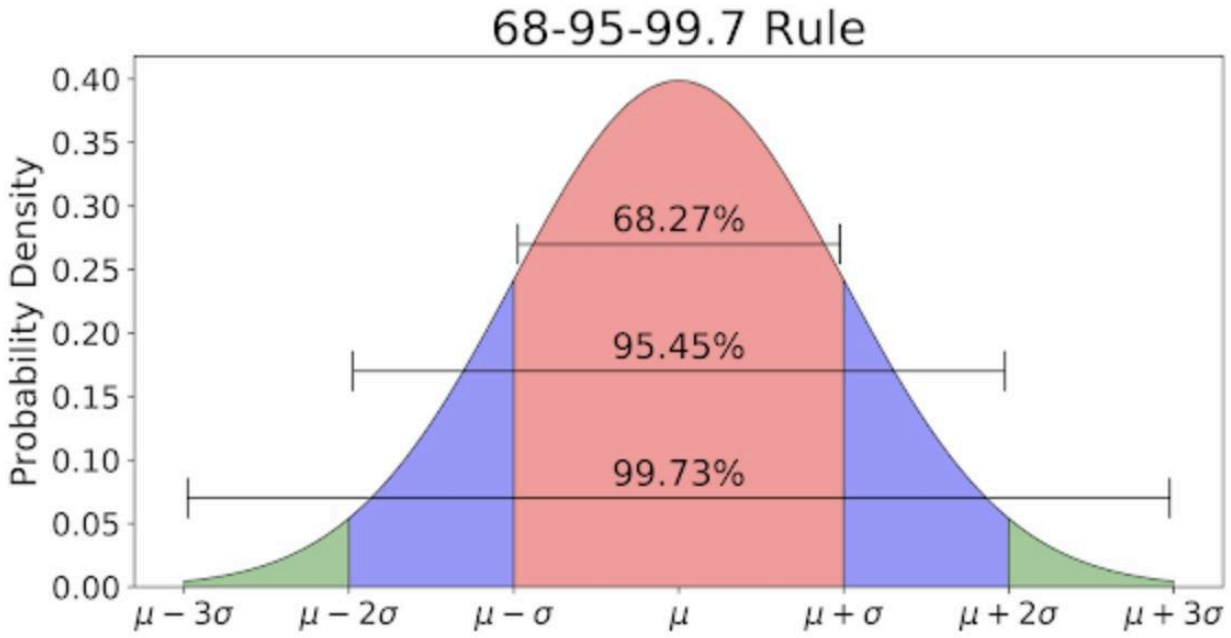
Visual Representation:

Key Points:

- ***Applies to Normal Distributions:*** *The Empirical Rule is specifically applicable to data that follows a normal distribution (bell-shaped curve).*
- ***Quick Estimation:*** *It provides a quick way to estimate the proportion of data that falls within certain ranges around the mean.*
- ***Limitations:*** *It's an approximation and may not be perfectly accurate for all normal distributions.*

In Summary:

The Empirical Rule is a fundamental concept in statistics that helps us understand the distribution of data in a normal distribution.[9] By knowing the mean and standard deviation, we can quickly estimate the percentage of data that falls within different ranges around the mean.



40. What is the relationship between sample size and power in hypothesis testing?(66)

In hypothesis testing, the relationship between sample size and power in hypothesis testing is ***direct and positive.*** *A larger sample size gives greater statistical power, which means that it's more likely to avoid a Type II error (false negative).*

So, ***Power -*** *The probability of correctly rejecting the null hypothesis when it is actually false and* ***Sample Size -*** *The number of observations included in the study are positively related because Larger samples provide more accurate estimates of the population parameters. This reduces the variability and uncertainty associated with the sample, leading to more precise results.(****Reduced Sampling Error****)*

*And With a larger sample, you're more likely to detect a true difference or effect if it exists. This is because larger samples provide more statistical evidence to support or reject the null hypothesis(****Better Detection of True Effects****) and hence larger sample sizes has higher power*

In some cases, it's more accurate to study a well-selected sample than to study the entire population. However, conducting a study of the entire population may provide more accurate findings.

41. Can you perform hypothesis testing with non-parametric methods?(67)

Yes, you can absolutely perform hypothesis testing with non-parametric method and are used when Parametric assumptions are violated. When your data doesn't fit the "perfect" bell curve

required by standard tests (like the t-test), you turn to ***non-parametric methods****. These are often called "distribution-free" tests because they don't assume your data follows a specific mathematical distribution like the Normal distribution.*

They are particularly useful in your cluster analysis if your data is skewed (like Income) or consists of rankings rather than exact measurements.

1. When to Use Non-Parametric Methods?

You should choose these methods if:

- ***The distribution is skewed:*** *Your data has long tails or outliers.*
- ***Small sample sizes:*** *You have too few data points to prove normality.*
- ***Ordinal Data:*** *Your data is ranked (e.g., "Satisfied," "Neutral," "Dissatisfied").*
- ***Inequality of Variance:*** *The groups you are comparing have very different spreads.*

2. Common Non-Parametric Alternatives
Most parametric tests have a non-parametric "twin." Instead of comparing ***means*** *(averages), these tests usually compare* ***medians*** *or the* ***ranks*** *of the data.*

Parametric Test (Normal Data)	***Non-Parametric Twin (Any Data)***	***What it compares***
Independent t-test	***Mann-Whitney U Test***	*Differences between 2 independent groups.*
Paired t-test	***Wilcoxon Signed-Rank Test***	*Differences between 2 related groups (before/after).*
One-Way ANOVA	***Kruskal-Wallis Test***	*Differences between 3 or more groups.*
Pearson Correlation (r)	***Spearman's Rank (ρ)***	*Strength of relationship between variables.*

How it Works: The Logic of Ranking

*Non-parametric tests work by stripping away the "raw values" and replacing them with **ranks**.*

***Example:** If you compare spending between Cluster A and Cluster B:*

1. *Combine all spending values from both groups.*
2. *Order them from smallest to largest.*
3. *Assign a rank (1, 2, 3...) to each.*
4. *If Cluster A's ranks are consistently higher than Cluster B's, the test concludes there is a significant difference.*

Finally, the following table can help you understand when and where you should use the parametric tests or their non-parametric counterparts and their advantages and disadvantages.

ANALYTIX LABS

Criterion	Parametric	Non Parametric
Population	A proper understanding of the population is available	Not much information about the population is available
Assumptions	Several assumptions are regarding the population. Incorrect results are provided if assumptions are not fulfilled	No assumptions are made regarding the population
Distribution	The distribution of the population is often required to be normal	Do not require the population to be normal; it can be arbitrary
Sample Size	Require sample size to be over 30	Can work with small samples
Interpretability	Are easy to interpret	Are difficult to interpret
Implementation	Are difficult to implement	Are easy to implement
Reliability	The output is more powerful/reliable	Are less powerful/reliable
Type of variable	Works with continuous/ quantitative variables	Works with continuous/quantitative as well as categorical/discreet variables
Central Tendency	Measurement of the central tendency is typically done using mean	Measurement of the central tendency is generally done using median
Outliers	Affected by outliers	Less affected by outliers
Null Hypothesis	More accurate	Incorrectly rejects null hypothesis; Less accurate
Examples	z-test, t-test, ANOVA, f-test, Pearson coefficient of correlation	One sample KS test, Wilcoxon signed rank test, Mann-Whitney U-test, Wilcoxon rank-sum test, Wilcoxon signed-rank test, Kruskal-Wallis test, Spearman's rank correlation, Kuiper's test, Hosmer-Lemeshow test, Chi-Square test for independence

---

4. Pros and Cons

The Advantages

- ***Robustness:*** *They aren't "tricked" by extreme outliers (like that $666,666 income we found earlier).*
- ***Versatility:*** *They work on almost any type of data scale.*

The Disadvantages

- ***Less Power:*** *If your data is actually normal, non-parametric tests are less likely to find a significant result (higher risk of Type II error). To compensate for this 'less power,' you need to increase the sample size to gain the result that the parametric counterpart would have provided.*
- *We are rarely interested in a significance test alone; we would like to say something about the population from which the samples came, and this is best done with estimates of parameters and confidence intervals.*
- *It is difficult to do flexible modelling with non-parametric tests, for example allowing for confounding factors using multiple regression.*
- *Parametric tests usually have more statistical power than their non-parametric equivalents. In other words, one is more likely to detect significant differences when they truly exist.*

So, the test to be used depends on what types of data are being measured as the test used should be determined by the data.

42. What factors affect the width of a confidence interval?(68)

The width of a confidence interval is affected by sample size, confidence level, and standard deviation-

- *Increasing your sample size is the primary way to reduce the widths of confidence intervals because, in most cases, you can control it more than the variability.*
- *If you increase the confidence level (e.g., 95% to 99%) while holding the sample size and variability constant, the confidence interval widens.*
- *Variability present in your data affects the precision of the estimate. Your confidence intervals will be broader when your sample standard deviation is high.*

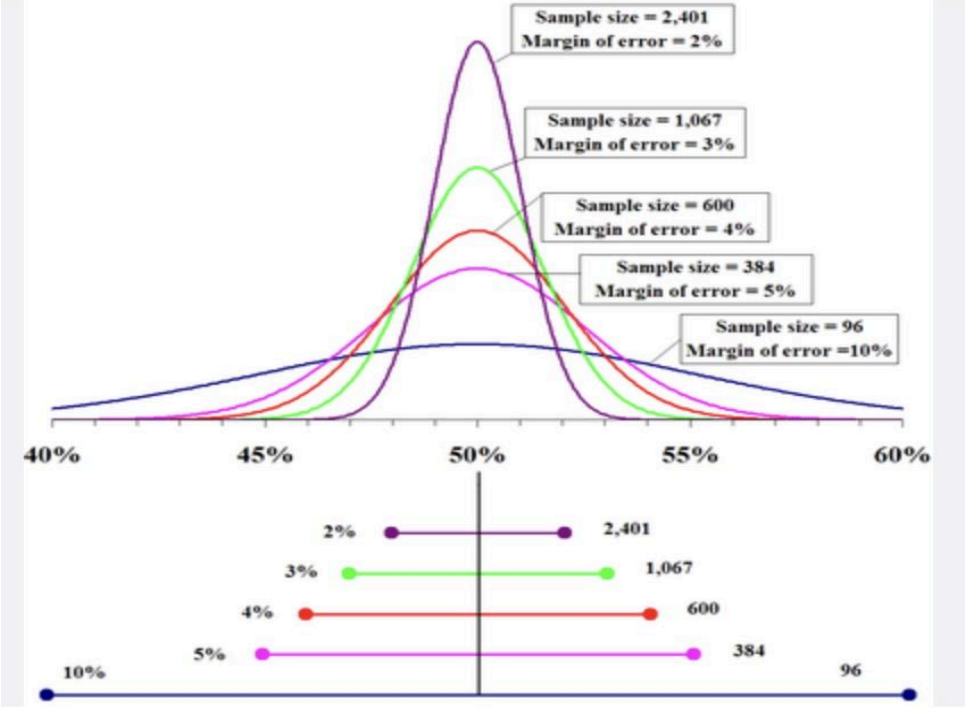
The confidence level also affects the confidence interval width. However, this factor is a methodology choice separate from your sample's characteristics.

If you increase the confidence level (e.g., 95% to 99%) while holding the sample size and variability constant, the confidence interval widens. Conversely, decreasing the confidence level (e.g., 95% to 90%) narrows the range.

Imagine you take your knowledge of a subject area and indicate you're 95% confident that the correct answer lies between 15 and 20. Then I ask you to give me your confidence for it falling between 17 and 18. The correct answer is less likely to fall within the narrower interval, so your confidence naturally decreases.

Conversely, I ask you about your confidence that it's between 10 and 30. That's a much wider range, and the correct value is more likely to be in it. Consequently, your confidence grows.

Confidence levels involve a tradeoff between confidence and the interval's spread. To have more confidence that the parameter falls within the interval, you must widen the interval. Conversely, your confidence necessarily decreases if you use a narrower range.



43. **Can a confidence interval be used to make a definitive statement about a specific individual in the population?(70)**

No, a confidence interval cannot be used to make a definitive statement about a specific individual in the population. ***This is because***

- ***Confidence intervals focus on population parameters:*** *Confidence intervals are designed to estimate population parameters, such as the population mean or proportion. They provide a range of plausible values for these parameters based on the sample data.*
- ***Individuals are not the focus:*** *They don't directly address the characteristics of individual members within the population.*

Example: *Imagine you have a bag of marbles. A confidence interval for the average weight of the marbles tells you the range within which the average weight of all marbles in the bag likely falls. It doesn't tell you the weight of any specific marble in the bag.*
While confidence intervals provide valuable information about the population as a whole, they don't offer insights into the specific attributes of individual members of that population.

44. **What is the relationship between the margin of error and confidence interval?(72)**
Confidence Interval:

- *A confidence interval is a range of values within which we are fairly certain (with a certain level of confidence) that the true population parameter lies.*
- *It's expressed as a range, such as "between X and Y."*

Margin of Error:

- *The margin of error is the amount by which the sample statistic (like the sample mean) might differ from the true population parameter.*
- *It's the "plus or minus" value that's added and subtracted from the sample statistic to create the confidence interval.*

FORMULA:

Confidence Interval = Sample Statistic ± Margin of Error

The precise formula depends on the type of parameter you're evaluating. You'll use critical Z-values or t-values to calculate your confidence interval of the mean.

Where:

- $\bar{x}$ *= the sample mean, which is the point estimate.*
- Z *= the critical z-value*
- t *= the critical t-value*
- s *= the sample standard deviation*
- $s / \sqrt{n}$ *= the standard error of the mean*

To calculate a confidence interval, take the critical value (Z or t) and multiply it by the standard error of the mean (SEM). This value is known as the margin of error (MOE). Then add and subtract the MOE from the sample mean ($\bar{x}$) to produce the upper and lower limits of the range.

45. What is a Chi-Square test?(75)

The Chi-square is a non-parametric test. It is a hypothesis test that answers questions like—do the values of one categorical variable depend on the value of other categorical variables, do the observed data fits a particular distribution ?.They cannot have a normal distribution because they have only a few particular values.There are two main types of Chi-Square tests:

Key Uses:

- *1.* ***Chi-Square Test for Independence*** *(This test is also known as the chi-square test of association.)*
  Example: A researcher wants to determine if there is an association between gender (male/female) and preference for a new product (like/dislike). The test can assess whether preferences are independent of gender.
- *2. Chi-Square Test for Goodness of Fit:Used to determine if observed data fits a particular distribution (like a normal distribution or a uniform distribution)-uniform distribution in this case.*
  Example: A dice manufacturer wants to test if a six-sided die is fair. They roll the die 60 times and expect each face to appear 10 times. The test checks if the observed frequencies match the expected frequencies.

How it Works:

1.Let's say you want to know if gender has anything to do with political party preference. You poll 440 voters in a simple random sample to determine their preferred political party. The results of the survey are shown in the table below:

	Republican	Democrat	Independent	Total
Male	100	70	30	200
Female	140	60	20	220
Total	240	130	50	440

To see if gender is linked to political party preference, perform a Chi-Square test of independence using the steps below.

Step 1: *Define the Hypothesis*

H0: There is no link between gender and political party preference.

H1: There is a link between gender and political party preference.

Step 2: *Calculate the Expected Values*

Now, you will calculate the expected frequency.

$$\text{Expected Value} = \frac{(Row\ Total) * (Column\ Total)}{Total\ Number\ Of\ Observations}$$

For example, the expected value for Male Republicans is:

$$= \frac{(240) * (200)}{440} = 109$$

Similarly, you can calculate the expected value for each of the cells.

Expected Values				
	Republican	**Democrat**	**Independent**	**Total**
Male	109	59	22.72	200
Female	120	65	25	220
Total	240	130	50	440

Step 3: *Calculate (O-E)2 / E for Each Cell in the Table*

Now, you will calculate the (O - E)2 / E for each cell in the table.

where

O = Observed Value

E = Expected Value

$(O - E)^2/E$				
	Republican	**Democrat**	**Independent**	**Total**
Male	0.74311927	2.050847	2.332676056	200
Female	3.33333333	0.384615	1	220
Total	240	130	50	440

Step 4: *Calculate the Test Statistic X2*

X2 is the sum of all the values in the last table

= 0.743 + 2.05 + 2.33 + 3.33 + 0.384 + 1

= 9.837

Before you can conclude, you must determine the critical statistic, which requires determining our degrees of freedom. The degrees of freedom in this case are equal to the table's number of columns minus one multiplied by the table's number of rows minus one, or (r-1) (c-1). We have (3-1)(2-1) = 2.

Finally, you compare the obtained statistics to the critical ones in the chi-square table. As you can see, for an alpha level of 0.05 and two degrees of freedom, the critical statistic is 5.991, less than our obtained statistic of 9.83. You can reject our null hypothesis because the critical statistic is higher than your obtained statistic.

Note:

1. ***we can programmatically calculate the chi-square test values using scipy.stats module***
2. ***When you have smaller sample sizes, you might need to use Fisher's exact test instead of the chi-square version***

46. What is a t-test?(76)

The t-test is a parametric statistical analysis hypothesis test used to compare the means of two groups.

Types of t-tests:

One-sample t-test: *Compares the mean of a single group to a known or hypothesized value.*

Independent samples t-test: *Compares the means of two independent groups (e.g., treatment group vs. control group).*

Paired samples t-test: *Compares the means of two related groups (e.g., the same individuals*

measured before and after a treatment).

Key Assumptions:

Normality: *The data in each group should be approximately normally distributed.*

Independence: *Observations within each group should be independent of each other.*

Equal Variances (for independent samples t-test): *The variances of the two groups should be equal (homoscedasticity).*

How it works:

Formulate Hypotheses:

Null Hypothesis (H0): *There is no significant difference between the means of the two groups.*

Alternative Hypothesis (H1): *There is a significant difference between the means of the two groups.*

Calculate the t-statistic: *This statistic measures the difference between the group means relative to the variability within the groups.*

Determine Degrees of Freedom: *The degrees of freedom depend on the sample sizes of the groups.*

Find the p-value: *The p-value represents the probability of obtaining the observed results (or more extreme results) if the null hypothesis were true.*

Make a Decision:

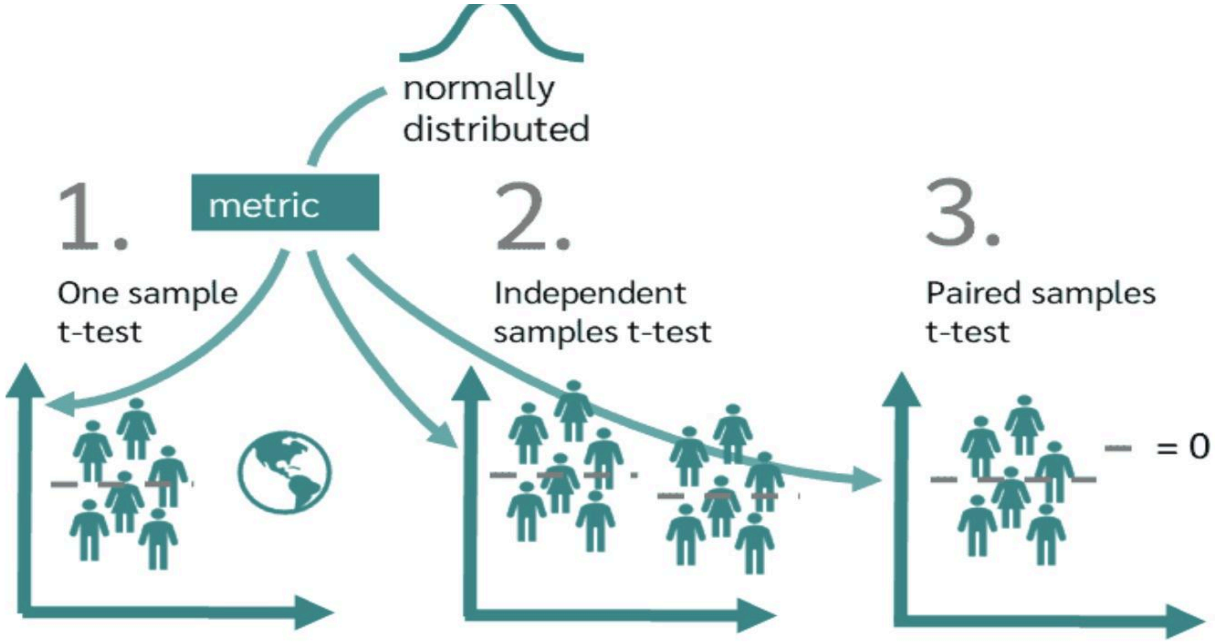
If the p-value is less than the chosen significance level (usually 0.05), reject the null hypothesis. This suggests that there is a statistically significant difference between the means of the two groups.

If the p-value is greater than or equal to the significance level, fail to reject the null hypothesis.

Assumptions:

1. *You have a random sample*
2. *A t test requires continuous data.*
3. *Your sample data follow a normal distribution, or you have a large sample size*
4. *Population standard deviation is unknown*

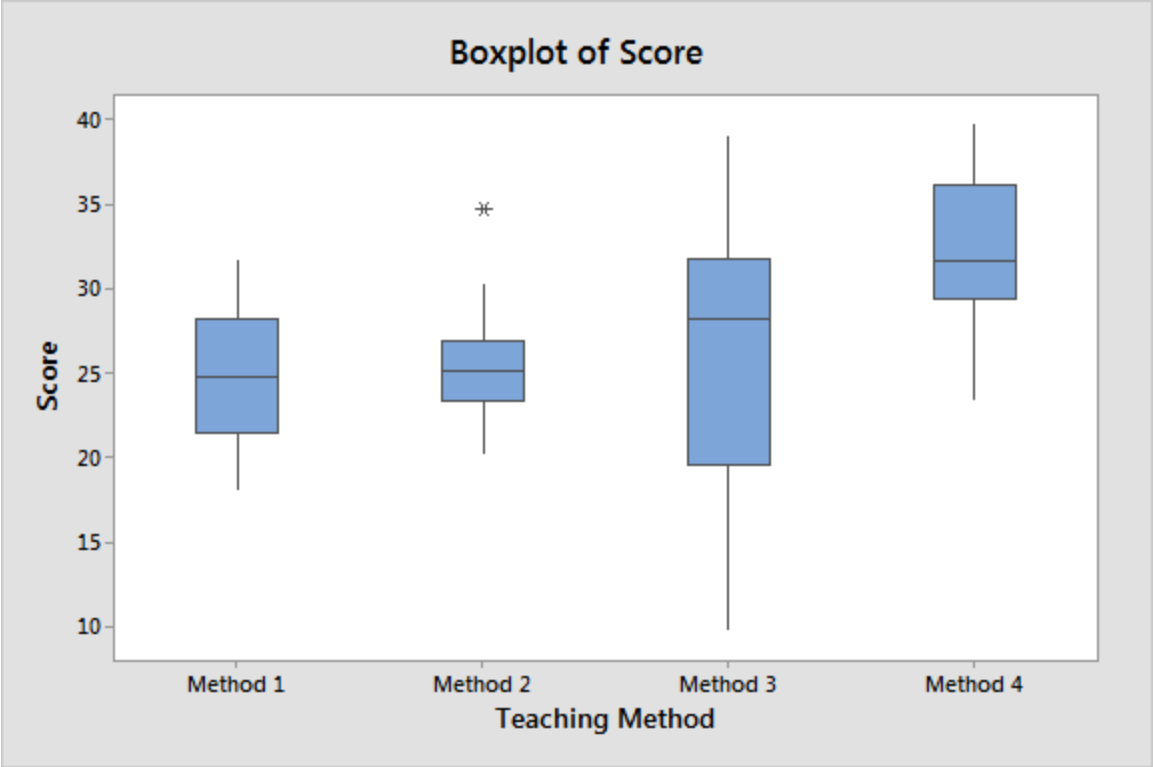
In summary: *The t-test is a valuable statistical tool for comparing means and making inferences about the differences between groups.It's important to understand the assumptions of the t-test and choose the appropriate type of t-test based on the research question and the characteristics of the data.*

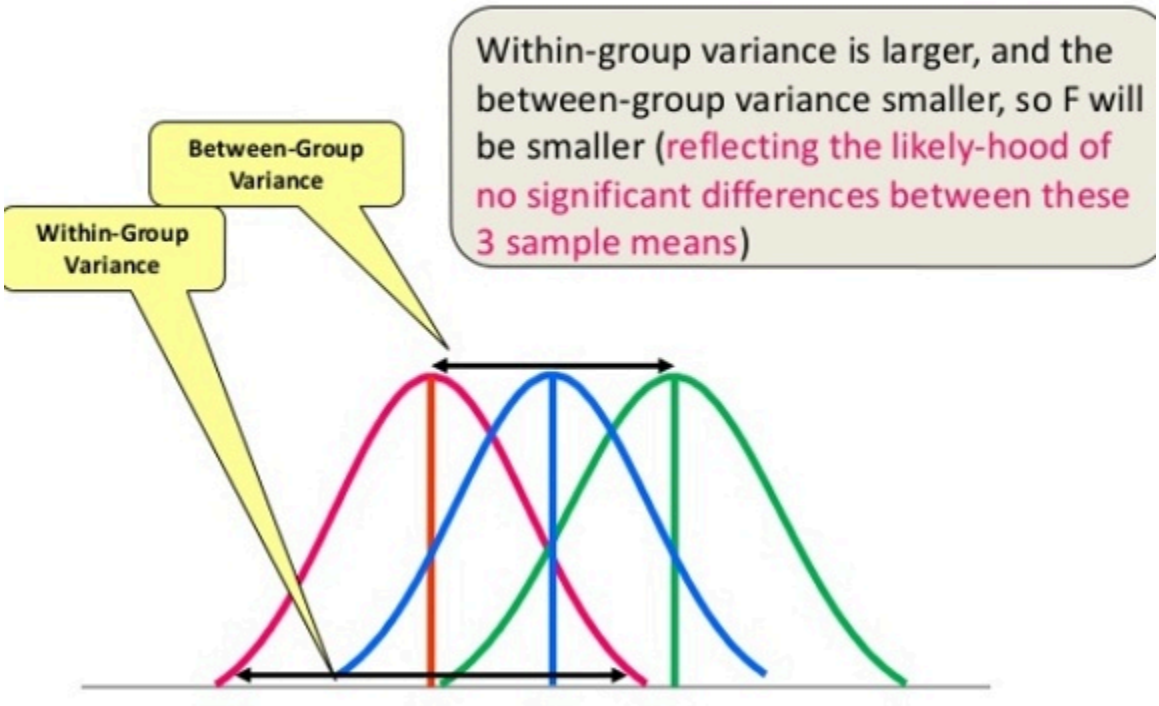


47. What is the ANOVA test?(77)

Analysis of variance (ANOVA) compares means across groups, assessing if observed differences are statistically significant, using variances within and between groups. ***ANOVA analyzes the variance within each group and the variance between groups. If the variance between groups is significantly larger than the variance within groups, it suggests that the group means are likely different.*** *In Exploratory Data Analysis (EDA), ANOVA helps identify significant variations between groups, aiding insights into data patterns.*

Another measure to compare the samples is called a t-test. When we have only two samples, t-test, and ANOVA give the same results. However, using a t-test would not be reliable in cases with more than 2 samples.


Boxplot of Score
Score
40
35
30
25
20
15
10
Method 1
Method 2
Method 3
Method 4
Teaching Method


Between-Group Variance
Within-Group Variance
Within-group variance is larger, and the between-group variance smaller, so F will be smaller (reflecting the likely-hood of no significant differences between these 3 sample means)

This process determines if the groups are part of one larger population or separate populations with different means. Consequently, even though it analyzes variances, it actually tests means

***Assumptions**:*

ANOVA tests have the same assumptions as other linear models other than requiring a factor. Specifically:

- *The dependent variable is continuous.*
- *You have at least one categorical independent variable (factor).*
- *The observations are independent.*
- *The groups should have roughly equal variances (scatter).*
- *The data in the groups should follow a normal distribution.*
- *The residuals satisfy the ordinary least squares assumptions.*

One-Way ANOVA

***Purpose:** To determine if there are statistically significant differences between the means of three or more independent groups.*

Assumptions:

- ***Normality:** Data within each group should be approximately normally distributed.*
- ***Homogeneity of variances:** The variance within each group should be equal (homoscedasticity).*
- ***Independence:** Observations within each group should be independent of each other.*

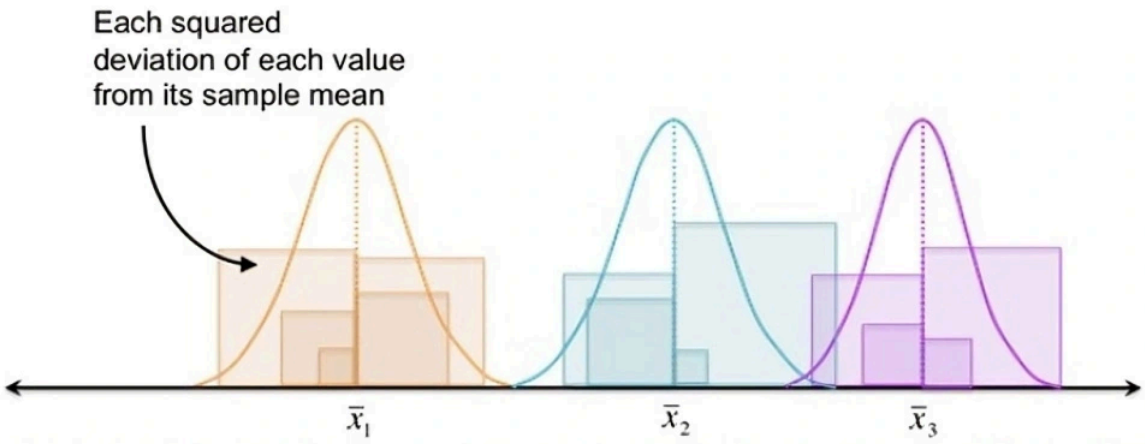
Steps:

1. ***State the Hypotheses:***
   - ***Null Hypothesis (H0):** $\mu 1 = \mu 2 = \mu 3 = ... = \mu k$ (All group means are equal)*
   - ***Alternative Hypothesis (H1):** At least one group mean is different from the others.*
   - ***Calculate the F-statistic:***
     - ***F-statistic = MSB / MSW***

Key Formulae:

1. ***Within-Groups Sum of Squares (SSW):***

Note: X_{i1} is the ith value from the first sample, X_{iz} is the ith value from the second sample, and so on all the way to X_{ik} the ith value from the kth sample. X_{ij} is therefore the ith value from the jth sample.

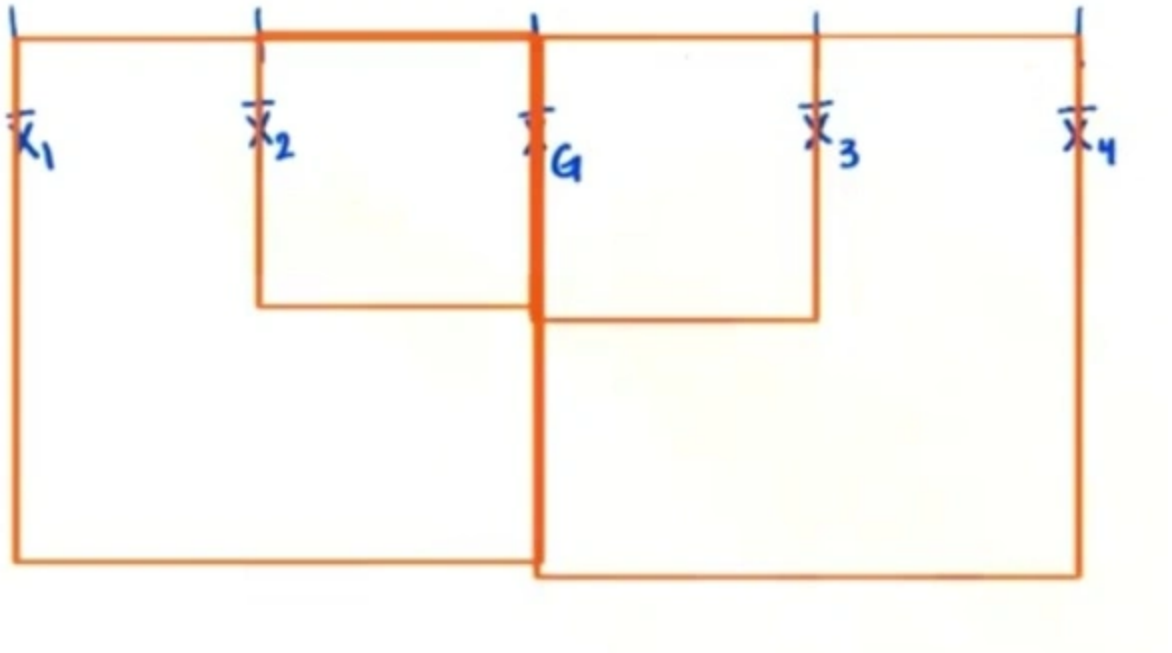


with within-group variability, SS_{within} is the sum of each squared deviation of each value from its respective sample mean (the total area of all the squares in the figure above). MS_{within} is the average-sized square.

- *For each group, calculate the mean (**$\bar{y}i$**).*
- *For each observation within a group, calculate (x_ij - **$\bar{y}i$**)^2.*
- *Sum these values for all observations across all groups.*

2. Calculate Between-Groups Sum of Squares (SSB):

$$SS\_between = n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2 + n_3(\bar{y}_3 - \bar{y})^2 + \ldots + n\square(\bar{y}\square - \bar{y})^2$$



Then divide by the degrees of freedom (), which in the case of between-group variability is the number of sample means (k) minus 1.

3. Calculate Total Sum of Squares(SST)

The Total Sum of Squares (SST or TSS) measures the total variability in a dataset by summing the squared differences between each observed data point (y_i) and the overall sample mean (ȳ). It acts as a measure of dispersion, where SST = sum (yi - ȳ)^2, and is used in regression and ANOVA to partition variance.

$$SST = \Sigma(y_i - \bar{y})^2$$

1. y_i = *Observed values*
2. $\bar{y}$ = *Mean of the observed values*

Then check whether **SST (Total Sum of Squares),** *accounts for total variation from overall mean.*

$$SST = SSB + SSW$$

$$SST = SSR + SSE$$

4. Calculate Degrees of Freedom:

- ***dfbetween = k - 1*** *(where k is the number of groups)*
- ***dfwithin = N - k*** *(where N is the total number of observations)*
- ***dftotal = dfbetween + dfwithin***

- ***Calculate Mean Squares:***
  - ***Mean Square Between (MSB) = SSB / dfbetween***
  - ***Mean Square Within (MSW) = SSW / dfwithin***

4. Calculate the F-statistic:

- ***F-statistic = MSB / MSW***

- ***Determine the p-value:***

Use an F-distribution table or statistical software to find the p-value associated with the calculated referring to the F-Distribution table, using ***dfbetween*** *and* ***dfwithin*** *at given significance level say α = 0.05 which is* ***Ftable***

Make a Decision:

- *If the p-value is less than the chosen significance level (usually 0.05), reject the null hypothesis ie If* ***Ftable< F-statistic*** *=> Reject null hypothesis*

This indicates that there are statistically significant differences between the means of at least two groups.

- *If the p-value is greater than or equal to the significance level, fail to reject the null hypothesis ie* ***Ftable>F-statistic => Fail to Reject null hypothesis***

. There is no sufficient evidence to conclude that the group means are different.

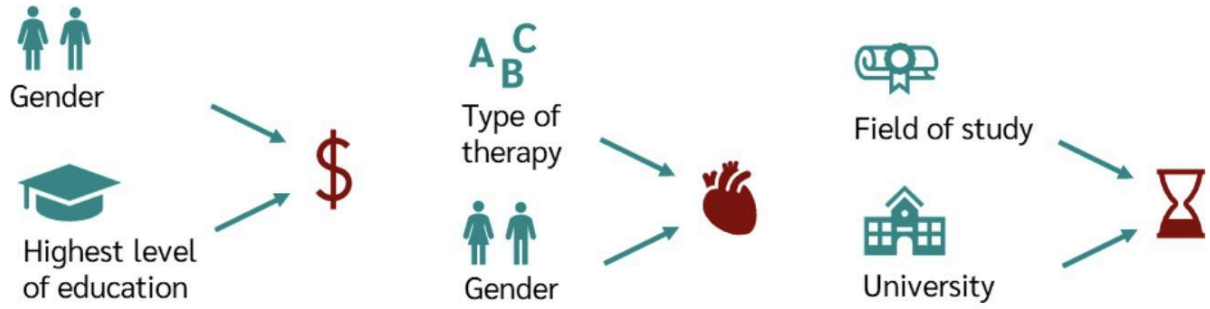
Example : *Assume it is necessary to assess whether consuming a specific type of tea will result in a mean weight decrease. Allow three groups to use three different varieties of tea: green tea, Earl Grey tea, and Jasmine tea. Here, the ANOVA test (one way) will be utilized to examine if there was any mean weight decrease displayed by a certain group.*

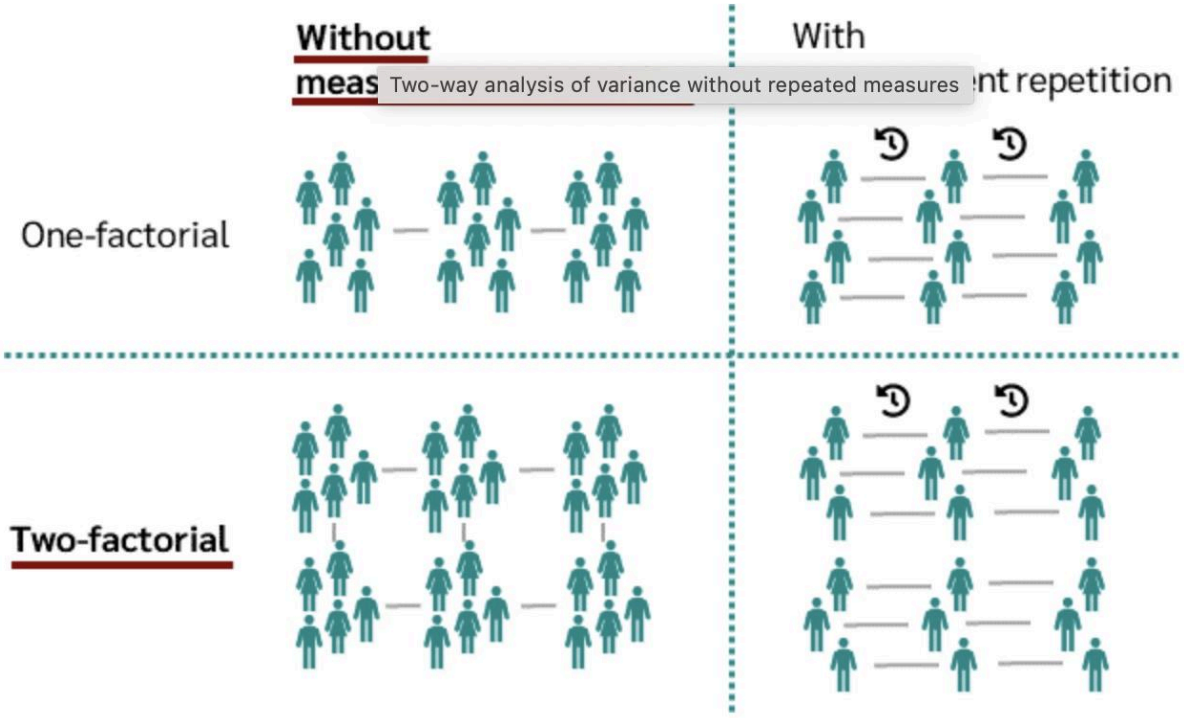
Post-Hoc Tests: *If the ANOVA test indicates significant differences between the group means, further analysis (such as post-hoc tests like Tukey's HSD or Bonferroni) is typically conducted to determine which specific groups differ from each other.*

Types of ANOVA:

1. ***One-way ANOVA:*** *Compares the means of three or more independent groups.*
2. ***Two-way ANOVA:*** *Compares the means of groups based on two independent variables (factors).*
3. ***Repeated Measures ANOVA:*** *Used when the same individuals are measured multiple times under different conditions.*
4. ***Multivariate analysis of variance (MANOVA)***

In Summary: *ANOVA is a powerful statistical tool that allows researchers to compare the means of multiple groups simultaneously. It provides valuable insights into whether observed differences between groups are meaningful or simply due to random chance.*
When we do ANOVA, WE ALWAYS DO A RIGHT TAIL TEST





48. Imagine that Jeremy took part in an examination. The test has a mean score of 160, and it has a standard deviation of 15. If Jeremy's z-score is 1.20, what would be his score on the test?

Understand Z-Score

- *A z-score represents how many standard deviations a data point is away from the mean.*
- *A positive z-score means the data point is above the mean.*[1]
- *A negative z-score means the data point is below the mean.*[2]

Formula:
*Jeremy's Score = (Z-score * Standard Deviation) + Mean*
*= (1.20 * 15) + 160. —> Plug in the values*

=> Jeremy's Score = 18 + 160 = 178

49. In an observation, there is a high correlation between the time a person sleeps and the amount of productive work he does. What can be inferred from this?(62)

A high correlation between sleep duration and work productivity suggests a strong relationship between the two. However, it's crucial to remember that ***correlation does not equal causation****.*

Possible Inferences:

- ***Sufficient Sleep May Enhance Productivity:*** *Adequate sleep is vital for cognitive function, alertness, and overall well-being. Sufficient sleep can improve focus, concentration, and decision-making, all of which contribute to increased productivity.*
- ***Lack of Sleep May Impair Productivity:*** *Insufficient sleep can lead to fatigue, decreased alertness, and difficulty concentrating, all of which can negatively impact work performance.*

What We Cannot Infer: *While the correlation is strong, it doesn't definitively prove that increased sleep causes increased productivity. There could be other factors at play. We can't say for certain whether more sleep directly leads to better work,* ***or if other factors like stress, diet, or exercise also influence both sleep and productivity.***

Further Research Needed: *To establish a causal relationship between sleep duration and work productivity, further research would be needed, such as controlled experiments where sleep duration is manipulated while other factors are held constant.*

In Summary: *While the high correlation suggests a strong link between sleep and productivity, further investigation is necessary to determine the exact nature of this relationship and whether increased sleep directly causes increased productivity.*

50. If there is a 30 percent probability that you will see a supercar in any 20-minute time interval, what is the probability that you see at least one supercar in the period of an hour (60 minutes)?(64)

Let's define the event "S" as seeing a supercar in a 20-minute interval.

Probability of Not Seeing a Supercar in 20 Minutes: *If the probability of seeing a supercar in 20 minutes is 30% (0.3), then the probability of not seeing a supercar in 20 minutes is:*

P(not S) = 1 - P(S) = 1 - 0.3 = 0.7

Divide the Hour into Intervals- *Since there are 60 minutes in an hour and we're considering 20-minute intervals, there are 3 intervals in an hour.*

Probability of Not Seeing a Supercar in an Hour *- To not see a supercar in the entire hour, you must not see it in any of the three 20-minute intervals.*

P(not seeing a supercar in an hour) = P(not S) * P(not S) * P(not S) = 0.7 * 0.7 * 0.7 = 0.343

Probability of Seeing at Least One Supercar in an Hour - *The probability of seeing at least one supercar in an hour is the complement of not seeing any supercars:*

P(seeing at least one supercar) = 1 - P(not seeing any supercars)

P(seeing at least one supercar) = 1 - 0.343 = 0.657

Therefore, the probability of seeing at least one supercar in the period of an hour is 65.7%.

<u>Conclusion:</u>

The Applied Statistics Interview Grind showcased the critical role of statistical knowledge in problem-solving. It emphasized the need for a strong grasp of statistical methods to analyze and interpret data effectively. The interview sessions highlighted the versatility of statistical tools across various industries, emphasizing their application in decision-making processes. The emphasis on real-world problem-solving underscored the importance of statistical techniques in driving evidence-based insights and informed decision-making. Overall, the grind provided a glimpse into how statistical expertise forms the backbone of addressing complex challenges across diverse fields.