## **Bilal Hameed**

## **Data Scientist @ CareCloud**

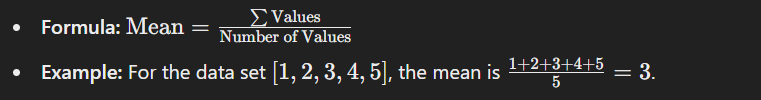
## **Machine Learning Interview Notes**

------------------------------------------------------------------------------------------------------------------------------------------

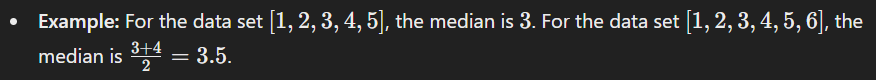
## **Explain the difference between mean, median, and mode. When would you use each?**

The mean, median, and mode are three different measures of central tendency used in statistics to describe the center point of a data set.

## **Mean**

* **Definition:** The mean (or average) is the sum of all the values in a data set divided by the number of values.
* **Use Case:** Best used when you want an overall average and the data does not have extreme outliers.

## **Median**

* **Definition:** The median is the middle value in a data set when the values are arranged in ascending or descending order. If there is an even number of values, the median is the average of the two middle values.
* **Calculation Steps:**
  1. Arrange the data in order.
  2. Find the middle value.
* **Use Case:** Best used when you need the central point of data and want to minimize the effect of outliers.

## **Mode**

* **Definition:** The mode is the value that appears most frequently in a data set. A data set may have one mode, more than one mode, or no mode at all.
* **Example:** For the data set [1,2,2,3,4], the mode is 2 because it appears most frequently. For the data set [1,1,2,3,3], the modes are 1 and 3 (bimodal).
* **Use Case:** Best used when you want to know the most common value(s) in the data set.

## **Summary**

* **Mean:** Sum of values divided by the number of values.
* **Median:** Middle value when data is ordered.
* **Mode:** Most frequent value(s) in the data set.

Each measure provides different insights and is useful in different scenarios depending on the nature of the data and the specific requirements of the analysis.

------------------------------------------------------------------------------------------------------------------------------------------

## **What is the difference between Descriptive and Inferential Statistics?**

**Descriptive Statistics** summarizes and describes the main features of a data set. It includes measures like the mean, median, mode, and standard deviation, as well as visual tools like graphs and charts. It’s all about presenting what the data shows without making any conclusions beyond the data itself.

**Inferential Statistics** goes a step further by using data from a sample to make predictions or inferences about a larger population. It involves techniques like hypothesis testing, confidence intervals, and regression analysis, helping you make decisions or predictions based on the data.

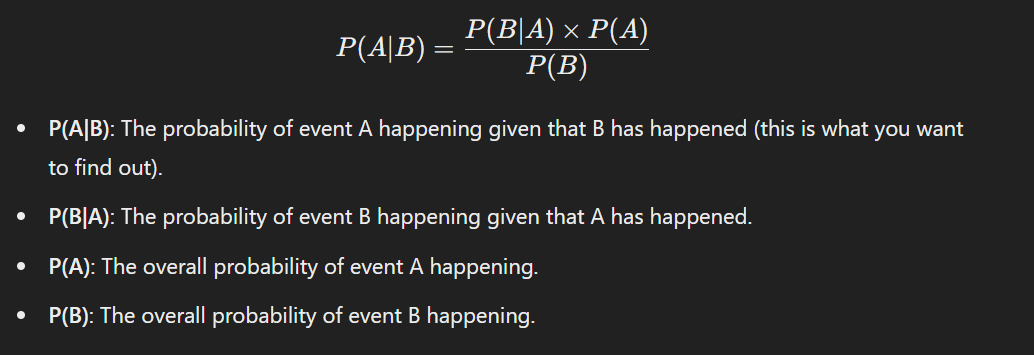
------------------------------------------------------------------------------------------------------------------------------------------

# **What is Bayes theorem and what is it about?**

Bayes' theorem is a mathematical formula that describes the probability of an event, based on prior knowledge of conditions that might be related to the event. The theorem is fundamental in probability theory and statistics, often used for updating probabilities as new evidence becomes available.

**OR**

Bayes' Theorem is a powerful tool in probability that helps you update your beliefs or predictions based on new evidence. It's a way of figuring out the probability of something happening, given that something else has already happened.

 ---------------------------------------------------------------------------------------------------------------

Bayes' Theorem is a way of finding a probability when we know certain other probabilities. It helps us update our beliefs about something based on new evidence. Here’s a simple breakdown:

1. **Prior Probability**: This is what you initially believe the probability of an event is before any new evidence. For example, if you believe there’s a 30% chance it will rain tomorrow, that’s your prior probability.
2. **Likelihood**: This is how likely the new evidence is, assuming your initial belief (prior) is correct. For instance, if you see dark clouds in the sky and think they are 80% likely to appear if it’s going to rain, that's your likelihood.
3. **Posterior Probability**: This is the updated probability after taking the new evidence into account. So after seeing the dark clouds, you might revise your belief about the chance of rain upwards.

In mathematical terms, Bayes’ Theorem combines these ideas as:

Posterior Probability = (Likelihood×Prior Probability)

Probability of the Evidence ​

Or simply, it tells you how to update your initial guess based on new information.

**---------------------------------------------------------------------------------------------------------------------**

## **What is Difference B/W probabilities and conditional probabilities?**

Probabilities and conditional probabilities are both concepts used in probability theory to quantify uncertainty, but they serve slightly different purposes.

1. **Probabilities (Marginal Probabilities):**
   * Probabilities, represent the likelihood of an event occurring without any conditions or additional information.
   * For example, if you roll a fair six-sided die, the probability of rolling a 3 is 1/6. This is a simple probability because it doesn't depend on any other information.
2. **Conditional Probabilities:**
   * Conditional probabilities, on the other hand, represent the likelihood of an event occurring given that another event has already occurred.
   * Mathematically, the conditional probability of event A given event B is denoted as P(A|B), and it's calculated as the probability of both events A and B happening divided by the probability of event B happening: P(A ∩ B) / P(B).
   * For example, if you roll two fair six-sided dice, and you know that the sum of the rolls is 7, then the conditional probability of the first die showing a 3 is 1/6, because there is only one outcome (3,4) out of the total outcomes (1,6), (2,5), (3,4), (4,3), (5,2), and (6,1) that satisfy the condition.

In summary, probabilities give the likelihood of an event occurring in general, while conditional probabilities give the likelihood of an event occurring given that some other event has already occurred. Conditional probabilities allow us to adjust our probabilities based on additional information or conditions.Top of Form

**---------------------------------------------------------------------------------------------------------------------**

## **What is probability Distribution?**

Probability distributions describe how the values of a random variable are spread out.

A probability distribution is a mathematical function or a table that describes the likelihood of different outcomes in a sample space. In simpler terms, it tells you how probable different values or events are in a given scenario. Probability distributions are fundamental in statistics and probability theory because they provide a way to model uncertainty and randomness in various phenomena.

There are two main types of probability distributions:

**Discrete Probability Distribution:**

* **Values**: A discrete probability distribution deals with variables that can take on a **countable number of distinct values**. These values are often whole numbers.
* **Example**: Rolling a die. The result can only be 1, 2, 3, 4, 5, or 6—no other values are possible.
* **Probability**: The probability is assigned to each individual value. For instance, the probability of rolling a 3 on a fair die is 1/6​.

**Continuous Probability Distribution:**

* **Values**: A continuous probability distribution deals with variables that can take on **any value within a range**. These values can be whole numbers or fractions, and they can have infinitely many possibilities within the range.
* **Example**: The exact height of people. A person's height could be 170.1 cm, 170.15 cm, or 170.155 cm, and so on—there's no limit to the precision.
* **Probability**: The probability of any single exact value is technically zero because there are infinitely many possible values. Instead, probabilities are assigned to ranges of values. For example, the probability that someone's height is between 170 cm and 175 cm.

Probability distributions can be described using various parameters such as mean, variance, standard deviation, and shape parameters. These parameters provide insights into the central tendency, spread, and shape of the distribution.

Probability distributions play a crucial role in statistical analysis, hypothesis testing, modeling of real-world phenomena, and decision-making under uncertainty. They allow researchers and analysts to make predictions, estimate probabilities, and draw conclusions based on available data.

**------------------------------------------------------------------------------------------------------------------------------------------**

## **What is diff between normally distributed vs multinomial distributions?**

Normally distributed and multinomial distributions are two different types of probability distributions, each with its own characteristics and applications:

1. **Normally Distributed (Gaussian) Distribution:**
   * The normal distribution, also known as the Gaussian distribution, is a continuous probability distribution that is symmetric around its mean.
   * It is characterized by two parameters: the mean (μ) and the standard deviation (σ).
   * In a normal distribution, the data tends to cluster around the mean, with the probability decreasing as you move away from the mean.
   * The famous bell-shaped curve represents the normal distribution, and it is widely used in statistics due to its properties, such as the central limit theorem.
   * Examples of naturally occurring phenomena that can be modeled with a normal distribution include heights of people, errors in measurements, and test scores.
2. **Multinomial Distribution:**
   * The multinomial distribution is a generalization of the binomial distribution to more than two categories.
   * It describes the probability of observing counts within each of multiple categories, where each observation falls into exactly one category.
   * Unlike the normal distribution, the multinomial distribution is discrete rather than continuous.
   * It is characterized by the number of categories (n) and a vector of probabilities (p₁, p₂, ..., pₙ) representing the probabilities of each category.
   * Examples of situations modeled by a multinomial distribution include outcomes of rolling a fair six-sided die, results of an election with multiple candidates, or outcomes of drawing colored balls from an urn with replacement.

In summary, the main differences between normally distributed and multinomial distributions lie in their form (continuous vs. discrete), the parameters they are characterized by (mean and standard deviation vs. number of categories and probabilities), and the types of data they model (observations around a mean vs. counts in multiple categories).

**---------------------------------------------------------------------------------------------------------------------**

## **Difference B/W parametric method Vs non-parametric method.**

### **1. Parametric Methods**

* **What Are They?**  
  These methods assume that the data follows a certain distribution (like a normal distribution, which is bell-shaped).
* **Key Feature:**  
  They have a fixed number of parameters. For example, a normal distribution is defined by just two parameters: mean and standard deviation.
* **Example:**  
  Suppose you want to estimate the average height of people in a city. If you assume the heights are normally distributed, you're using a parametric method.
* **Pros:**
  + If the assumption about the distribution is correct, these methods can be very powerful and accurate.
  + They generally require less data to get good results.
* **Cons:**
  + If your assumption about the data's distribution is wrong, the method might give misleading results.

### **2. Non-Parametric Methods**

* **What Are They?**  
  These methods do **not** assume any specific distribution for the data. Instead, they are more flexible and can adapt to different shapes and patterns in the data.
* **Key Feature:**  
  They don't have a fixed number of parameters. The model's complexity can grow with more data.
* **Example:**  
  If you want to estimate the average height of people without assuming the data is normally distributed, you might just look at the data directly or use a method like the median, which is non-parametric.
* **Pros:**
  + More flexible since they don’t assume a specific distribution.
  + Useful when you have little idea about the underlying distribution of the data.
* **Cons:**
  + Usually require more data to get accurate results.
  + Can be computationally intensive and harder to interpret.

### **When to Use Which?**

* **Parametric:** When you have a good reason to believe that your data follows a specific distribution.
* **Non-Parametric:** When you don’t want to make strong assumptions about the data's distribution or when the data doesn't fit any known distribution well.

In simple terms, parametric methods are like using a recipe to bake a cake because you assume you know the ingredients. Non-parametric methods are like experimenting with different ingredients because you're not sure what the recipe should be.

**---------------------------------------------------------------------------------------------------------------------**

## **What is Random Variables in statistic?**

In statistics, a **random variable** is a variable whose possible values are numerical outcomes of a random phenomenon. It essentially represents a quantity that can take different values depending on the outcome of a random event.

### **Key Concepts of Random Variables:**

1. **Types of Random Variables**:
   * **Discrete Random Variable**: Can take on a finite or countably infinite number of possible values. Examples include the roll of a die (which can result in one of six values) or the number of heads in a series of coin flips.
   * **Continuous Random Variable**: Can take on an infinite number of possible values within a given range. Examples include the exact height of individuals or the time taken for a task, where the values can vary continuously within a range.
2. **Probability Distribution**:
   * Each random variable is associated with a probability distribution, which specifies the likelihood of each possible outcome. For a discrete random variable, this is known as a **probability mass function (PMF)**, and for a continuous random variable, it's called a **probability density function (PDF)**.
3. **Expected Value (Mean)**:
   * The expected value of a random variable is a measure of the central tendency, or average, of the possible outcomes. It’s calculated differently for discrete and continuous random variables but represents the long-term average if the random process were repeated many times.
4. **Variance and Standard Deviation**:
   * These measures describe the spread or variability of the random variable’s possible outcomes around the mean. The variance is the average of the squared differences from the mean, and the standard deviation is the square root of the variance.

### Example:

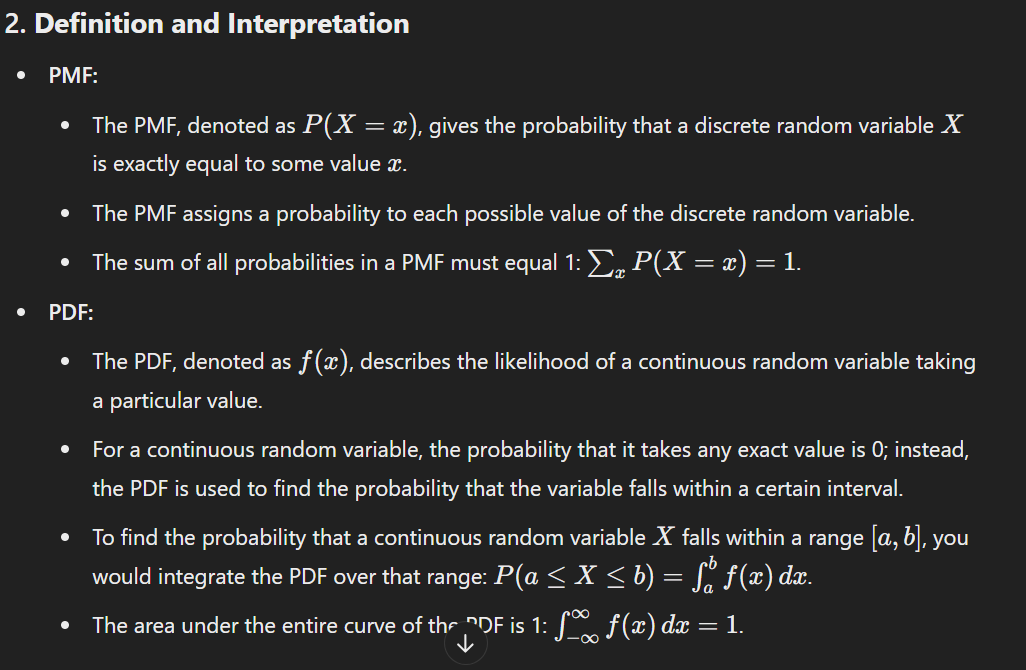
* **Discrete Random Variable Example**: Suppose we have a random variable X that represents the outcome of rolling a six-sided die. X can take any value from 1 to 6, each with an equal probability of 1/6
* **Continuous Random Variable Example**: Suppose we have a random variable Y that represents the height of a randomly chosen person. Y could take any value within a reasonable range, say between 4.5 and 7 feet, with different probabilities associated with different ranges of heights.

In summary, random variables are foundational concepts in probability and statistics, enabling us to model and analyze uncertain or random processes mathematically.

**---------------------------------------------------------------------------------------------------------------------**

## **Diff Between probability mass function (PMF) and probability density function (PDF).**

The concepts of Probability Mass Function (PMF) and Probability Density Function (PDF) are fundamental in probability theory and statistics, and they are used to describe the distribution of discrete and continuous random variables, respectively. Here's a breakdown of the key differences:



**PMF** is used for discrete variables, assigning a probability to each distinct outcome.

**PDF** is used for continuous variables, describing the probability density over a continuous range of outcomes.

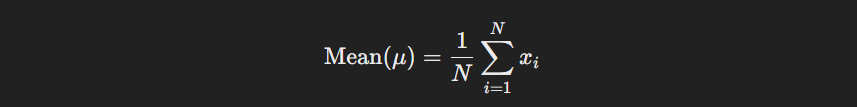
**---------------------------------------------------------------------------------------------------------------------**

## **How do you calculate the variance and standard deviation of a dataset?**

To calculate the variance and standard deviation of a dataset, follow these steps:

### 1. **Calculate the Mean (Average):**

* Sum all the data points.
* Divide the sum by the number of data points.



Where *N* is the number of data points and *xi​* represents each data point.

### 2. **Calculate Each Deviation from the Mean:**

* Subtract the mean from each data point to find the deviation of each point from the mean.

Deviation=*xi−μ*

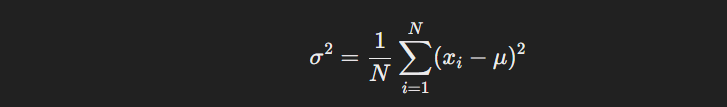
### 3. **Square Each Deviation:**

* Square each of the deviations calculated in step 2 to remove negative values.

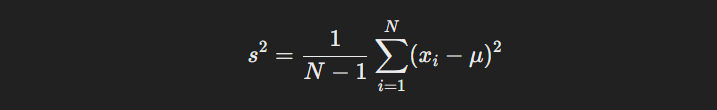
(Deviation)^2=(xi−μ)^2

### 4. **Calculate the Variance:**

* Find the average of these squared deviations.
* For a **population variance** (if you're considering the data as the entire population):



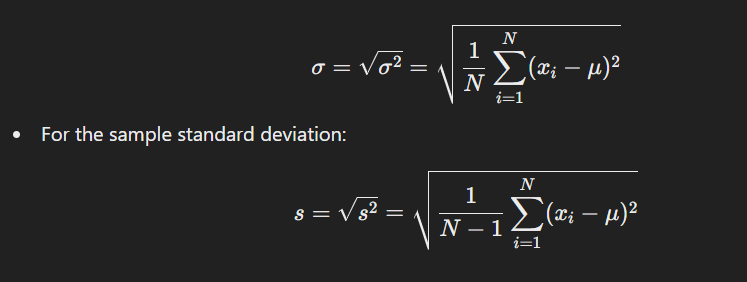
* For a **sample variance** (if you're considering the data as a sample of a larger population):



The sample variance uses N−1 in the denominator (known as Bessel's correction) to correct for the bias that occurs when estimating a population parameter from a sample.

### 5. **Calculate the Standard Deviation:**

* Take the square root of the variance to get the standard deviation.
* For the population standard deviation:



### Summary

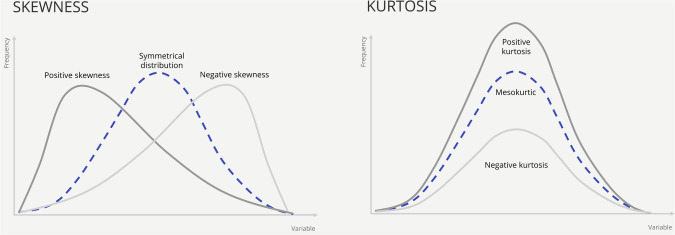
* **Variance** gives a measure of how spread out the data points are around the mean.
* **Standard deviation** is the square root of the variance and provides a measure of the average distance of each data point from the mean, in the same units as the data.

These calculations help in understanding the dispersion or variability within a dataset.

---------------------------------------------------------------------------------------------------------------------

**What is skewness and kurtosis? How do they help in understanding data distribution?**

**Skewness** and **kurtosis** are two statistical terms that help us understand the shape of a data distribution. Let me explain each in simple terms:



## **1. Skewness:**

Skewness measures the lack of symmetry in a data distribution. It indicates that there are significant differences between the mean, the mode, and the median of data. Skewed data cannot be used to create a normal distribution.

* **Skewness** tells us about the asymmetry or tilt in a data distribution.
* Imagine you have a graph showing how your data is spread out. If the graph looks perfectly symmetrical (like a bell curve), the skewness is zero.
* If the graph is **tilted to the right** (with a longer tail on the right), it's called **positive skew**. This means there are **more low values and fewer high values**.
* If the graph is **tilted to the left** (with a longer tail on the left), it's called **negative skew**. This means there are **more high values and fewer low values.**

## **Kurtosis:**

Kurtosis is used to describe the extreme values present in one tail of distribution versus the other. It is actually the measure of outliers present in the distribution. A high value of kurtosis represents large amounts of outliers being present in data. To overcome this, we have to either add more data into the dataset or remove the outliers.

* **Kurtosis** tells us about the **tailedness or how heavy the tails** of the distribution are.
* If the distribution has **normal tails**, it's called **mesokurtic** (like a normal distribution).
* If the distribution has **heavy tails** (more extreme values), it's called **leptokurtic/Positive Kurtosis**. This means more data points are in the tails or center, leading to a peakier graph.
* If the distribution has **light tails** (fewer extreme values), it's called **platykurtic/ Negative Kurtosis**. This means the graph is flatter, with fewer data points in the tails or center.

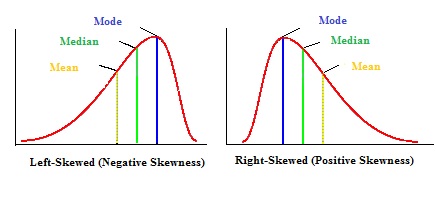
### How Do They Help?

* **Skewness** helps you see if your data is leaning more towards higher or lower values, which can be important in decision-making. For example, if you're analyzing incomes and the data is positively skewed, most people earn less, with a few earning a lot.
* **Kurtosis** helps you understand the presence of outliers (extreme values). High kurtosis indicates many outliers, while low kurtosis suggests fewer outliers.

Together, skewness and kurtosis give you a more detailed picture of your data, beyond just knowing the average or range. They help in understanding the shape and spread of the data, which is crucial in many areas like finance, research, and quality control.

---------------------------------------------------------------------------------------------------------------------

## **What are left-skewed and right-skewed distributions?**



A left-skewed distribution is one where the left tail is longer than that of the right tail. Here, it is important to note that the **mean < median < mode.**

Similarly, a right-skewed distribution is one where the right tail is longer than the left one. But, here **mean > median > mode.**  
---------------------------------------------------------------------------------------------------------------------

**What is the central limit theorem, and why is it important in statistics?**

The Central Limit Theorem (CLT) is one of the most important concepts in statistics. Here’s an easy way to understand it:

### What is the Central Limit Theorem?

The Central Limit Theorem says that if you take a large number of random samples from any population, no matter what the shape of that population is (whether it’s skewed, uniform, or even weirdly shaped), the average of those samples will form a distribution that is approximately normal (bell-shaped).

### Breaking It Down:

1. **Population:** Imagine you have a population of things—this could be the heights of all people in a city, the daily temperatures in a month, or the number of cats in households across a country. The distribution of these values might be anything—it could be skewed, have multiple peaks, etc.
2. **Sampling:** Now, you start taking samples from this population. A sample is just a small group taken randomly from the population. Each time you take a sample, you calculate the average (mean) of that sample.
3. **Forming a Distribution:** After taking many samples and calculating the average for each, you plot these averages on a graph.
4. **The Magic:** The Central Limit Theorem tells us that no matter the shape of the original population's distribution, the distribution of these sample averages will start to look like a bell curve (normal distribution) if the sample size is large enough.

### Why Is This Important?

* **Prediction and Inference:** It allows us to make predictions about population parameters (like the mean) even if we don’t know the population’s exact distribution.
* **Confidence Intervals:** The CLT is the reason why we can create confidence intervals and conduct hypothesis tests, as these rely on the assumption of normality.

### Example to Illustrate:

Imagine you have a very weird-looking dice that’s not evenly weighted. If you roll it many times, you might get an odd-looking distribution of results (more threes, fewer ones, etc.). But if you roll it 30 times, take the average of those 30 rolls, and repeat this process over and over, the distribution of these averages will look like a normal bell curve.

### Key Points:

* **Sample Size Matters:** The theorem holds true when the sample size is large enough. Usually, a sample size of 30 is considered sufficient, but larger samples are better.
* **Any Distribution:** The original population’s distribution can be anything—normal, skewed, bimodal, etc.
* **Averages, Not Raw Data:** The CLT applies to the distribution of sample averages, not the raw data itself.

In summary, the Central Limit Theorem is a powerful concept that allows statisticians to use the normal distribution to make inferences about populations, even when we don’t know much about the population’s actual distribution.

---------------------------------------------------------------------------------------------------------------------

## **Describe different types of probability distributions (e.g., normal, binomial, Poisson).**

Probability distributions describe how the values of a random variable are spread out. They show the likelihood of different outcomes. Here’s a simple overview of some common types:



### 1. **Normal Distribution (Bell Curve)**

* **Shape**: Looks like a bell; it’s symmetrical.
* **Example**: Heights of people, test scores.
* **Key Points**: Most data points are around the average (mean), and fewer are far from it. If you plot it, the curve peaks in the middle and tapers off towards the ends.

### 2. **Binomial Distribution**

* **Shape**: Varies, but often looks like a series of bars or spikes.
* **Example**: Flipping a coin 10 times and counting how many heads you get.
* **Key Points**: It describes situations where there are two possible outcomes (like success/failure or yes/no) repeated a certain number of times. The distribution shows the probability of getting a certain number of successes.

### 3. **Poisson Distribution**

* **Shape**: Skewed, usually leaning to the right.
* **Example**: Number of emails you get per hour, number of cars passing by in a minute.
* **Key Points**: It models the number of times an event happens in a fixed interval of time or space, assuming these events happen independently and at a constant average rate.

### 4. **Uniform Distribution**

* **Shape**: A flat line.
* **Example**: Rolling a fair die (each number from 1 to 6 has an equal chance of showing up).
* **Key Points**: Every outcome has the same probability. The distribution is flat because all outcomes are equally likely.

### 5. **Exponential Distribution**

* **Shape**: Decreasing curve, starts high and then drops.
* **Example**: Time between arrivals of buses at a stop.
* **Key Points**: It’s often used to model the time between events that happen continuously and independently at a constant average rate.

### 6. **Geometric Distribution**

* **Shape**: Decreasing bars.
* **Example**: Number of coin flips until you get the first head.
* **Key Points**: It describes the number of trials needed to get the first success in a series of independent, identical trials.

### 7. **Bernoulli Distribution**

* **Shape**: Two bars (one at 0, one at 1).
* **Example**: Flipping a single coin.
* **Key Points**: It’s a special case of the binomial distribution with just one trial. It describes a single experiment where there are only two possible outcomes (success/failure).

These distributions help in modeling and understanding different kinds of real-world scenarios where randomness and uncertainty play a role.

---------------------------------------------------------------------------------------------------------------------

## **Explain the difference between a population and a sample. Why is sampling important?**

### Difference Between a Population and a Sample

* **Population**: A population includes all members or items that meet a particular criterion or set of criteria. It is the entire group about which you want to draw conclusions. For example, if you're studying the average height of adult men in a country, the population would include every adult man in that country.
* **Sample**: A sample is a subset of the population that is selected for the actual study or analysis. The sample should ideally represent the population from which it is drawn. For instance, if you can't measure the height of every adult man in the country, you might select a few thousand men as a sample to estimate the average height.

### Importance of Sampling

1. **Feasibility**: In many cases, it is impractical or impossible to study an entire population due to time, cost, or logistical constraints. Sampling allows researchers to gather and analyze data more efficiently.
2. **Cost-Effective**: Studying a smaller group (sample) instead of an entire population reduces the resources needed, making research more affordable.
3. **Time-Efficient**: Analyzing a sample is quicker than analyzing an entire population, which speeds up the research process and allows for faster decision-making.
4. **Manageability**: Working with a smaller sample makes data collection, storage, and analysis more manageable. Large data sets can be complex and challenging to work with.
5. **Accuracy**: If done correctly, sampling can provide accurate estimates of population parameters. Proper sampling techniques ensure that the sample represents the population, allowing for generalization of the results.
6. **Ethical Considerations**: In some cases, it may be unethical to study an entire population (e.g., testing a new drug on all patients). Sampling allows researchers to conduct studies in a way that minimizes harm or inconvenience.

In summary, while a population encompasses all members of a defined group, a sample is a manageable subset used for study. Sampling is crucial because it makes research feasible, cost-effective, and accurate while allowing researchers to draw meaningful conclusions about a larger group.

---------------------------------------------------------------------------------------------------------------------

## **How to Formulate Null and Alternative Hypotheses?**

1. **Identify the Research Question:**
   * Start with a clear research question or problem you want to investigate.
   * Determine what you want to prove or test.
2. **Formulate the Null Hypothesis (H₀):**
   * Assume no effect, no difference, or no change. This hypothesis should be stated in a way that can be tested directly.
   * Example: "The average weight loss from the new diet is equal to the average weight loss from the existing diet."
   * **H₀:** The population parameter equals a specific value (e.g., H0:μ=0 if testing for no difference).
3. **Formulate the Alternative Hypothesis (H₁ or Hₐ):**
   * This hypothesis should reflect the claim or effect you want to investigate. It is usually the opposite of the null hypothesis.
   * Example: "The average weight loss from the new diet is different from the average weight loss from the existing diet."
   * **H₁:** The population parameter differs from the value stated in H0​ (e.g., H1:μ≠0 for a two-tailed test).
4. **Determine the Type of Test (One-tailed or Two-tailed):**
   * **One-tailed test**: Used when you are testing for an effect in a specific direction (e.g., greater than or less than).
   * **Two-tailed test**: Used when you are testing for an effect in either direction (e.g., not equal).
5. **Set Up the Hypotheses:**
   * **Example:** Suppose a company claims that their light bulbs last for 1000 hours on average. You want to test this claim.
     + **Null Hypothesis (H₀):** The average life of the light bulbs is 1000 hours ( H0:μ=1000 ).
     + **Alternative Hypothesis (H₁):** The average life of the light bulbs is not 1000 hours ( H1:μ≠1000 ).

In summary, the null hypothesis is the claim you seek to test (often a statement of no effect), and the alternative hypothesis is what you are trying to find evidence for (a statement of an effect or difference).

---------------------------------------------------------------------------------------------------------------------\_\_\_\_\_\_\_

## **Diff B/w P value and critical Value?**

The **p-value** and the **critical value** are both concepts used in hypothesis testing, but they serve different roles and have different interpretations.

### **1. P-Value:**

* **Definition:** The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.
* **Interpretation:** It measures the strength of the evidence against the null hypothesis. A smaller p-value indicates stronger evidence against the null hypothesis.
* **Decision Rule:** Compare the p-value with the significance level (denoted as α\alphaα, often 0.05).
  + If p-value ≤α, reject the null hypothesis.
  + If p-value >α, fail to reject the null hypothesis.
* **Continuous Measure:** The p-value provides a continuous measure of evidence against the null hypothesis, so you get an exact value (e.g., 0.03) rather than a binary decision.

### **2. Critical Value:**

* **Definition:** The critical value is a threshold or cutoff point that defines the boundary for rejecting the null hypothesis. It is determined by the significance level (α\alphaα) and the distribution of the test statistic under the null hypothesis.
* **Interpretation:** It represents the value of the test statistic beyond which the null hypothesis would be rejected.
* **Decision Rule:** Compare the test statistic with the critical value.
  + If the test statistic is more extreme than the critical value, reject the null hypothesis.
  + If the test statistic is less extreme, fail to reject the null hypothesis.
* **Binary Decision:** The critical value approach gives a binary decision (reject or fail to reject the null hypothesis) based on whether the test statistic falls within the critical region.

### **Key Differences:**

* **Role:** The p-value is a measure of the evidence against the null hypothesis, while the critical value is a fixed threshold used to decide whether to reject the null hypothesis.
* **Comparison:** In p-value testing, you compare the p-value to α\alphaα. In critical value testing, you compare the test statistic to the critical value.
* **Flexibility:** The p-value allows for more flexibility and insight because it provides a continuous measure of significance, whereas the critical value provides a more straightforward, binary decision-making process.

In summary, both the p-value and critical value are tools used in hypothesis testing, but they offer different perspectives on making decisions about the null hypothesis.

------------------------------------------------------------------------------------------------------------------------------------------

## **Describe the steps in conducting a hypothesis test.**

Conducting a hypothesis test involves several key steps, which are designed to assess the evidence against a null hypothesis. Here’s a step-by-step guide to performing a hypothesis test:

### 1. **State the Hypotheses**

* **Null Hypothesis (H₀):** This is the default assumption that there is no effect or no difference. It represents the status quo or the claim to be tested.
* **Alternative Hypothesis (H₁ or Hₐ):** This is what you want to prove. It represents a new effect, difference, or relationship that contradicts the null hypothesis.

Example:

* H₀: The mean of the population is equal to 50.
* H₁: The mean of the population is not equal to 50.

### 2. **Choose the Significance Level (α)**

* The significance level, denoted as α, is the probability of rejecting the null hypothesis when it is actually true. Common choices for α are 0.05, 0.01, or 0.10, with 0.05 being the most common.

Example:

* Set α = 0.05.

### 3. **Select the Appropriate Test**

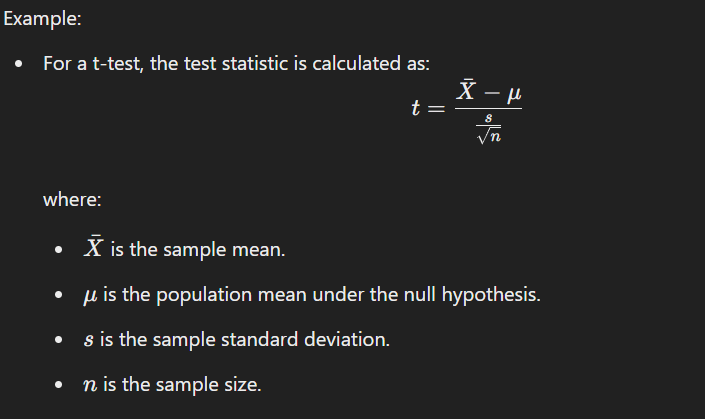
* Choose the statistical test that matches your data and hypothesis. The choice depends on factors like the type of data (e.g., categorical or continuous), sample size, and whether you are comparing means, proportions, etc.

Common tests include:

* **Z-test or t-test:** For comparing means.
* **Chi-square test:** For categorical data.
* **ANOVA:** For comparing means across multiple groups.
* **Regression analysis:** For relationships between variables.

### 4. **Calculate the Test Statistic**

* The test statistic in a hypothesis test is a value calculated from sample data that is used to determine whether to reject the null hypothesis. It measures how far the sample data deviates from what is expected under the null hypothesis. The test statistic helps quantify the difference between the observed sample data and the data expected under the null hypothesis. The formula and method for calculating the test statistic depend on the chosen test.



### 5. **Determine the p-value or Critical Value**

* **p-value:** In statistics, the p-value is a probability measure that helps you assess the evidence against a null hypothesis. It indicates the probability of observing the data or more extreme results, assuming the null hypothesis is true. A lower p-value suggests stronger evidence against the null hypothesis, often leading to its rejection in favor of an alternative hypothesis. Typically, a significance level (e.g., 0.05) is chosen, and if the p-value is below this threshold, the results are considered statistically significant.
* **Critical Value:** Alternatively, you can compare the test statistic to a critical value from a statistical distribution (e.g., t-distribution, Z-distribution) based on the chosen significance level.

Example:

* If using a p-value approach, calculate the p-value associated with the test statistic.
* If using a critical value approach, determine the critical value from statistical tables and compare it to the test statistic.

### 6. **Make a Decision**

* **Reject H₀:** If the p-value is less than or equal to α, or if the test statistic exceeds the critical value, reject the null hypothesis. This suggests there is sufficient evidence to support the alternative hypothesis.
* **Fail to Reject H₀:** If the p-value is greater than α, or if the test statistic does not exceed the critical value, fail to reject the null hypothesis. This suggests there is not enough evidence to support the alternative hypothesis.

### 7. **Draw a Conclusion**

* Based on the decision, conclude whether or not there is sufficient statistical evidence to support the alternative hypothesis. The conclusion should be stated in the context of the original research question or problem.

### 8. **Report the Results**

* Clearly report the test statistic, p-value, decision (whether H₀ was rejected), and the conclusion. It’s also helpful to include the confidence interval for the estimated effect size, if applicable.

Example:

* "The t-test results show that the mean is significantly different from 50 (t = 2.45, p = 0.017). Thus, we reject the null hypothesis at the 0.05 significance level.

Following these steps ensures a systematic and rigorous approach to hypothesis testing, allowing for sound conclusions based on statistical evidence.

---------------------------------------------------------------------------------------------------------------------

## **What is a p-value? How do you interpret it in the context of a hypothesis test?**

A **p-value** is a statistical measure that helps you determine the significance of your results in the context of a hypothesis test. It indicates the probability of obtaining a result at least as extreme as the one observed in your sample data, assuming that the null hypothesis is true.

### Interpreting the p-value in Hypothesis Testing:

1. **Null Hypothesis (H₀)**: This is the default assumption that there is no effect or no difference. For example, in a test comparing the means of two groups, the null hypothesis might state that the means are equal.
2. **Alternative Hypothesis (H₁)**: This represents the opposing assumption that there is an effect or a difference.
3. **p-value Meaning**:
   * **Low p-value (typically ≤ 0.05)**: This suggests that the observed data is unlikely under the null hypothesis. Therefore, you might reject the null hypothesis in favor of the alternative hypothesis. The lower the p-value, the stronger the evidence against the null hypothesis.
   * **High p-value (typically > 0.05)**: This suggests that the observed data is likely under the null hypothesis. Therefore, you fail to reject the null hypothesis. It does not necessarily prove that the null hypothesis is true, only that there isn't strong enough evidence to reject it.
4. **Significance Level (α)**:
   * The significance level, often denoted by α (e.g., 0.05), is a threshold set before conducting the test. If the p-value is less than or equal to α, you reject the null hypothesis; otherwise, you do not reject it.
   * For example, with α = 0.05, there is a 5% risk of rejecting the null hypothesis when it is actually true (Type I error).

### Example:

Imagine you are testing whether a new drug is more effective than a placebo. Your null hypothesis (H₀) might be that the drug has no effect (mean difference = 0). After conducting the test, you obtain a p-value of 0.03.

* Since 0.03 < 0.05 (assuming α = 0.05), you would reject the null hypothesis. This suggests that the drug likely has an effect, as the probability of observing such a result due to random chance is only 3%.

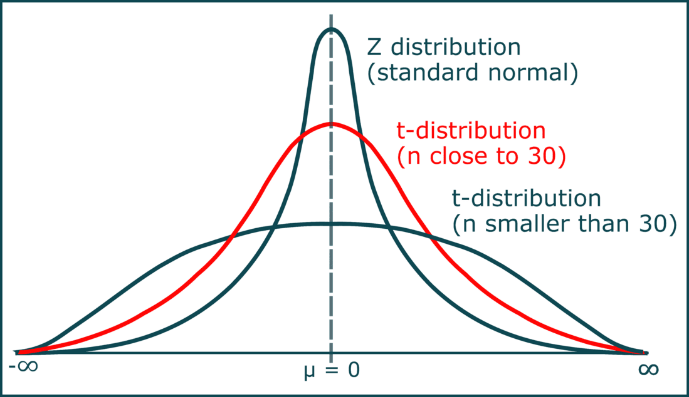
### Key Points:

* The p-value does not measure the probability that the null hypothesis is true.
* It also doesn't tell you the size of the effect or the importance of the result.
* It simply quantifies how compatible your data is with the null hypothesis.

---------------------------------------------------------------------------------------------------------------------

## **When would you use a t-test versus a z-test?**

Both t-tests and z-tests are statistical methods used to compare sample data to a known value or between two groups. The choice between a t-test and a z-test depends on several factors, including sample size, whether the population standard deviation is known, and the underlying distribution of the data.



### When to Use a T-Test

1. **Sample Size is Small (Typically n<30):**
   * When you have a small sample size, the t-distribution is used because it accounts for the added variability in the sample mean. The t-distribution is wider (has heavier tails) than the normal distribution, which helps account for the increased uncertainty.
2. **Population Standard Deviation is Unknown:**
   * The t-test is used when the population standard deviation (sigma σ) is unknown, and the sample standard deviation (sss) is used as an estimate. Since sss is just an estimate of σ sigma, the t-distribution, which is more conservative than the normal distribution, is appropriate.
3. **Small or Moderate Sample Size:**
   * Even with moderate sample sizes, if the population standard deviation is unknown, it's common practice to use a t-test, especially when n<30 n < 30n<30.

### When to Use a Z-Test

1. **Sample Size is Large (Typically n≥30n \geq 30n≥30):**
   * For large sample sizes, the Central Limit Theorem states that the sampling distribution of the sample mean tends to be normally distributed, regardless of the shape of the population distribution. In this case, you can use the z-test.
2. **Population Standard Deviation is Known:**
   * The z-test is used when the population standard deviation (sigma σ) is known. With this known value, the normal distribution (z-distribution) is appropriate, and the z-test is more straightforward.
3. **Comparing Proportions or Testing Hypotheses with Large Samples:**
   * Z-tests are also often used for hypothesis testing about proportions or for comparing means when sample sizes are large and population variances are known or assumed to be equal.

### Summary

* **Use a t-test** when:
  + Sample size is small.
  + Population standard deviation is unknown.
* **Use a z-test** when:
  + Sample size is large.
  + Population standard deviation is known.

In practice, with modern statistical software, the choice between t-test and z-test often defaults to a t-test because it handles more general cases and is appropriate even when the sample size is large (the t-distribution approximates the normal distribution closely as sample size increases).

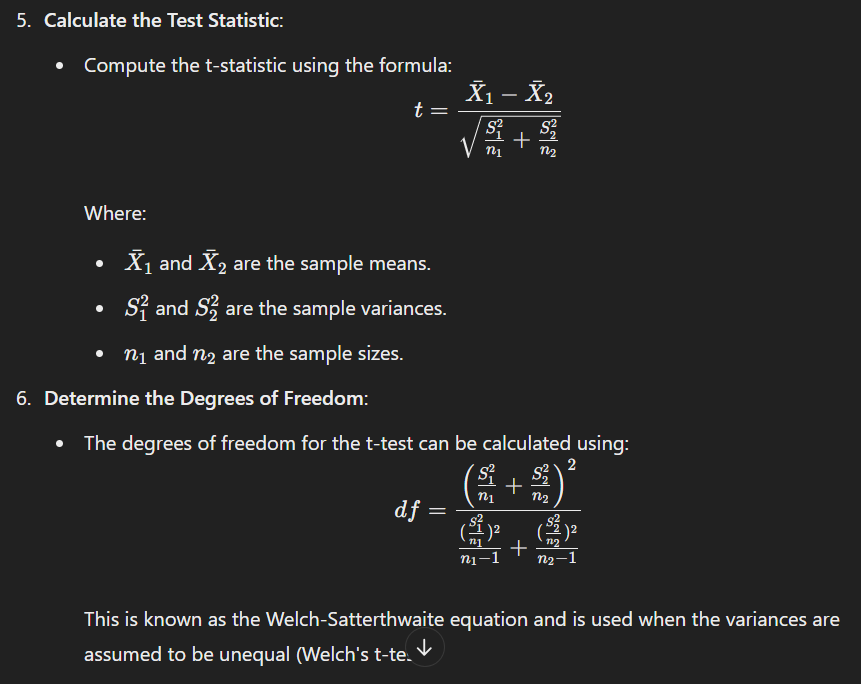
---------------------------------------------------------------------------------------------------------------------

## **Explain how you would conduct an independent two-sample t-test. What assumptions must be met?**

An independent two-sample t-test is used to compare the means of two independent groups to determine whether there is a statistically significant difference between them. Here's how you would conduct this test and the assumptions that need to be met:

### Steps to Conduct an Independent Two-Sample T-Test

1. **State the Hypotheses**:
   * **Null Hypothesis (H0​)**: The means of the two groups are equal (μ1=μ2 ​).
   * **Alternative Hypothesis (H1​)**: The means of the two groups are not equal (μ1≠μ2 ​).
2. **Choose the Significance Level (α\alphaα)**:
   * Common choices are 0.05, 0.01, or 0.10, depending on how strict you want to be.
3. **Collect the Data**:
   * Gather data for the two independent samples. Ensure the sample sizes are n1​ and n2​ for groups 1 and 2, respectively.
4. **Check Assumptions**:
   * Before proceeding with the t-test, verify that the data meet the necessary assumptions (detailed below).

****

1. **Find the Critical Value or P-Value**:
   * Compare the calculated t-statistic to the critical value from the t-distribution (based on the degrees of freedom) or compute the p-value.
   * If using a p-value approach, compare it to the significance level α alpha.
2. **Make a Decision**:
   * **If ∣t∣>tcritical ​ or p<α**: Reject the null hypothesis. There is a significant difference between the two means.
   * **If ∣t∣≤tcritical or p≥αp**: Fail to reject the null hypothesis. There is no significant difference between the two means.
3. **Report the Results**:
   * Present the findings, including the means, t-statistic, degrees of freedom, and p-value, and interpret them in the context of your research question.

### Assumptions of the Independent Two-Sample T-Test

1. **Independence of Observations**:
   * The observations in each group must be independent of each other. This means that the data from one group should not influence the data from the other group.
2. **Normality**:
   * The data in each group should be approximately normally distributed. This assumption is more critical when the sample size is small. For large sample sizes (usually n>30n > 30n>30 per group), the Central Limit Theorem ensures that the sampling distribution of the mean is approximately normal.
3. **Homogeneity of Variances (Equal Variances)**:
   * The variances of the two groups should be approximately equal. This assumption can be tested using Levene's test or an F-test for equality of variances.
   * If this assumption is violated, you should use Welch's t-test, which does not assume equal variances.
4. **Scale of Measurement**:
   * The dependent variable should be measured on an interval or ratio scale (continuous data).

By meeting these assumptions and following the steps outlined above, you can accurately perform an independent two-sample t-test and determine whether there is a significant difference between the two groups.

---------------------------------------------------------------------------------------------------------------------

## **Describe a scenario where you would use a paired sample t-test.**

Sure! A paired sample t-test is used when you have two sets of related measurements and you want to determine if there is a significant difference between the means of these two sets. Here’s a practical scenario where you might use a paired sample t-test:

**Scenario: Evaluating the Effectiveness of a New Diet Plan**

Imagine you’re a nutritionist conducting a study to evaluate the effectiveness of a new diet plan. You want to see if the diet plan leads to a significant reduction in body weight.

**Procedure:**

1. **Pre-Diet Measurement:** You measure the body weight of each participant before they start the diet plan. This gives you your first set of data.
2. **Post-Diet Measurement:** After a certain period on the diet plan (e.g., 8 weeks), you measure the body weight of the same participants again. This gives you your second set of data.

**Applying the Paired Sample t-Test:**

* **Null Hypothesis (H₀):** The average weight before starting the diet is equal to the average weight after following the diet (no change in weight).
* **Alternative Hypothesis (H₁):** The average weight before starting the diet is different from the average weight after following the diet (there is a significant change in weight).

You’d use a paired sample t-test to analyze the data and determine whether the observed changes in body weight are statistically significant, or if they could be due to random chance.

---------------------------------------------------------------------------------------------------------------------

## **What is ANOVA, and how does it differ from a t-test?**

ANOVA (Analysis of Variance) and the t-test are both statistical methods used to compare means, but they are applied in different contexts and have distinct purposes.

### ANOVA

**Purpose:** ANOVA is used to compare the means of three or more groups to determine if there is a significant difference between them. It tests the null hypothesis that all group means are equal.

**How It Works:** ANOVA works by partitioning the total variability in the data into variability due to the differences between group means and variability due to differences within each group. It then compares the ratio of these variabilities using an F-statistic to determine if the group means are significantly different.

**Types of ANOVA:**

* **One-way ANOVA:** Tests differences between group means for one independent variable.
* **Two-way ANOVA:** Tests differences between group means for two independent variables, and can also examine interaction effects between the variables.

### t-test

**Purpose:** The t-test is used to compare the means of two groups to determine if there is a significant difference between them.

**How It Works:** The t-test calculates a t-statistic that reflects the difference between the group means relative to the variability within the groups. It then uses this t-statistic to determine if the observed difference is statistically significant.

**Types of t-tests:**

* **Independent (or unpaired) t-test:** Compares the means of two independent groups.
* **Paired t-test:** Compares the means of two related groups (e.g., before and after treatment on the same subjects).

### Key Differences

1. **Number of Groups:**
   * **ANOVA:** Compares three or more groups.
   * **t-test:** Compares only two groups.
2. **Purpose:**
   * **ANOVA:** Tests if there are any significant differences among multiple group means.
   * **t-test:** Tests if there is a significant difference between the means of two groups.
3. **Outcome:**
   * **ANOVA:** Provides an F-statistic and p-value to indicate whether there is a significant difference among the groups.
   * **t-test:** Provides a t-statistic and p-value to indicate whether there is a significant difference between the two groups.

In summary, ANOVA is the go-to method when you need to compare means across more than two groups, while the t-test is appropriate for comparing the means of exactly two groups.

---------------------------------------------------------------------------------------------------------------------

## **What is the meaning of degrees of freedom (DF) in statistics?**

In statistics, "degrees of freedom" (DF) refers to the number of independent values or pieces of information in a calculation that are free to vary while still allowing us to estimate a particular statistic, like the mean or variance.

### **Simple Example:**

Imagine you have three numbers, and you know their average is 10. If you know the first two numbers (say, 8 and 12), the third number can't be anything — it must be 10 for the average to stay the same.

Here, out of the three numbers, two are "free" to be anything, but once you know them, the third is determined. This is why we say there are 2 degrees of freedom (DF = 2).

### **General Idea:**

* **Degrees of freedom** is the count of independent values that can vary in your dataset.
* The formula to calculate degrees of freedom usually depends on the context, but a common example is DF=n−1 when calculating the sample variance or standard deviation, where nnn is the number of data points. The subtraction by 1 accounts for the fact that once you know n−1 data points, the last one is not free to vary if you know the mean.

Understanding degrees of freedom is important in many statistical tests, as it affects how we calculate confidence intervals and significance levels.

---------------------------------------------------------------------------------------------------------------

## **Explain how you would interpret the results of a one-way ANOVA.**

Interpreting the results of a one-way ANOVA (Analysis of Variance) involves several key steps:

1. **Understand the Hypotheses:**
   * **Null Hypothesis (H₀):** The means of the different groups are equal (no effect of the treatment or factor).
   * **Alternative Hypothesis (H₁):** At least one group mean is different from the others.
2. **Examine the ANOVA Table:** The ANOVA table typically includes:
   * **Sum of Squares (SS):** Measures the variability in the data.
   * **Degrees of Freedom (df):** Reflects the number of independent pieces of information.
   * **Mean Squares (MS):** Calculated as SS divided by df.
   * **F-Statistic (F):** Ratio of the variance between groups to the variance within groups. It helps determine if the group means are significantly different.
   * **p-Value:** Probability of observing the data assuming the null hypothesis is true. A low p-value (typically less than 0.05) indicates that you can reject the null hypothesis.
3. **Check the F-Statistic:**
   * If the F-statistic is large, it suggests that there is a significant difference between the group means. This means the variability between group means is greater than the variability within groups.
4. **Interpret the p-Value:**
   * Compare the p-value to your significance level (α, usually 0.05). If the p-value is less than α, you reject the null hypothesis, concluding that there are significant differences between at least some group means. If the p-value is greater than α, you do not reject the null hypothesis, suggesting that there are no significant differences.
5. **Post-Hoc Tests (if necessary):**
   * If the ANOVA indicates significant differences, you may need to conduct post-hoc tests (like Tukey's HSD, Bonferroni, or Scheffé) to determine which specific groups are different from each other.
6. **Check Assumptions:**
   * **Normality:** The data in each group should be approximately normally distributed.
   * **Homogeneity of Variances:** The variance among the groups should be roughly equal.
   * **Independence:** The observations should be independent of each other.

By following these steps, you can determine whether there are significant differences between group means and identify which groups differ if necessary.

---------------------------------------------------------------------------------------------------------------------

## **Describe a situation where you might use a two-way ANOVA.**

A two-way ANOVA (Analysis of Variance) is useful when you want to examine the effect of two different categorical independent variables on a continuous dependent variable and also check if there’s an interaction between these two independent variables.

Here’s an example scenario:

**Research Question:** Does the type of diet and exercise regimen affect weight loss, and is there an interaction between these two factors?

**Design:**

* **Independent Variable 1 (Diet Type):** Two levels (e.g., Low-Carb and Low-Fat)
* **Independent Variable 2 (Exercise Regimen):** Two levels (e.g., Cardio and Strength Training)
* **Dependent Variable:** Weight loss (measured in pounds or kilograms)

**Setup:**

1. **Groups:** You have four groups based on the combination of diet and exercise regimen:
   * Low-Carb + Cardio
   * Low-Carb + Strength Training
   * Low-Fat + Cardio
   * Low-Fat + Strength Training
2. **Participants:** Assign a group of participants to each of the four groups. Over a period of time, measure the weight loss for each participant.

**Analysis with Two-Way ANOVA:**

* **Main Effects:** Determine if there are significant differences in weight loss due to diet type (Low-Carb vs. Low-Fat) and exercise regimen (Cardio vs. Strength Training).
* **Interaction Effect:** Investigate if the effect of one factor (e.g., diet) on weight loss depends on the level of the other factor (e.g., exercise regimen). For instance, the effect of the Low-Carb diet might be different depending on whether participants do Cardio or Strength Training.

By using a two-way ANOVA, you can not only assess the individual effects of diet and exercise but also understand if and how these factors interact to influence weight loss.

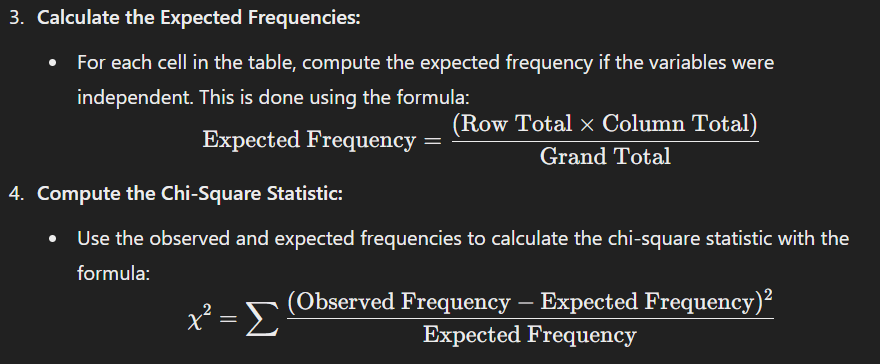
---------------------------------------------------------------------------------------------------------------------

## **What is a chi-square test for independence? When would you use it?**

A chi-square test for independence is a statistical method used to determine if there is a significant association between two categorical variables. In other words, it helps you figure out whether the distribution of one variable is independent of the distribution of another variable.

Here's how it works:

1. **Set Up the Hypotheses:**
   * **Null Hypothesis (H0):** The variables are independent; there is no association between them.
   * **Alternative Hypothesis (H1):** The variables are dependent; there is an association between them.
2. **Collect and Organize Data:**
   * Create a contingency table (cross-tabulation) that displays the frequency of occurrences for each combination of categories from the two variables.



1. **Determine the P-Value:**
   * Compare the chi-square statistic to a chi-square distribution with the appropriate degrees of freedom to find the p-value.
2. **Make a Decision:**
   * If the p-value is less than the chosen significance level (e.g., 0.05), reject the null hypothesis, suggesting there is a significant association between the variables. Otherwise, do not reject the null hypothesis.

**When to Use It:**

* **When you have two categorical variables:** For example, if you want to examine whether gender (male, female) is associated with voting preference (Democrat, Republican).
* **When you want to test the independence of variables in a sample:** For instance, checking if the distribution of educational levels is independent of employment status in a survey.
* **When your data is organized into a contingency table:** The chi-square test is appropriate when data can be summarized in this way, and you want to test for independence between the row and column variables.

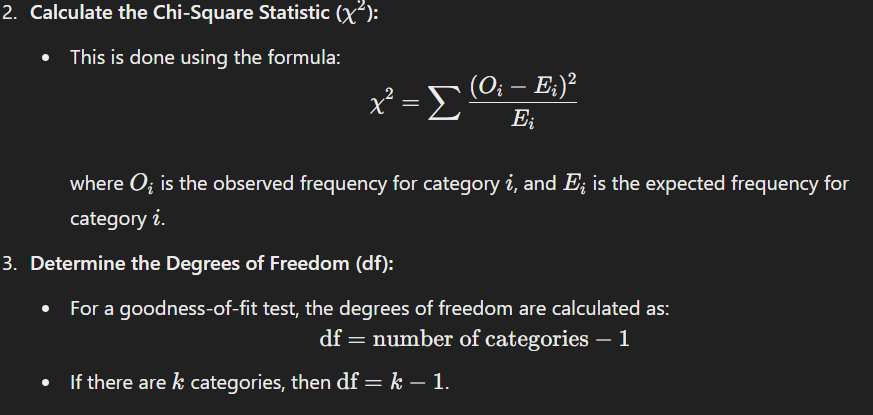
It’s important to ensure that the sample size is sufficiently large and that the expected frequency in each cell of the contingency table is adequate (generally at least 5) to make the chi-square test reliable. If this condition isn't met, you might need to use a different test, such as Fisher's exact test.

---------------------------------------------------------------------------------------------------------------------

## **How do you interpret the results of a chi-square goodness-of-fit test?**

Interpreting the results of a chi-square goodness-of-fit test involves several steps:

1. **Understand the Hypotheses:**
   * **Null Hypothesis (H0H\_0H0​)**: Assumes that the observed data fits the expected distribution. For example, if you’re testing whether a die is fair, H0H\_0H0​ would state that each face of the die has an equal probability of landing face up.
   * **Alternative Hypothesis (HAH\_AHA​)**: Assumes that the observed data does not fit the expected distribution.



1. **Find the P-value:**
   * Use the chi-square distribution table or statistical software to find the p-value corresponding to your calculated χ2\chi^2χ2 statistic and the degrees of freedom. The p-value tells you the probability of obtaining a chi-square statistic at least as extreme as the one calculated, assuming the null hypothesis is true.
2. **Compare the P-value to Your Significance Level (α\alphaα):**
   * Common significance levels are 0.05, 0.01, or 0.10.
   * If the p-value is less than or equal to α\alphaα, you reject the null hypothesis, suggesting that there is a significant difference between the observed and expected frequencies.
   * If the p-value is greater than α\alphaα, you do not reject the null hypothesis, suggesting that the observed frequencies do not significantly differ from the expected frequencies.
3. **Draw Conclusions:**
   * If you rejected the null hypothesis, it means the data does not fit the expected distribution. You may need to investigate further to understand the nature of the discrepancy.
   * If you did not reject the null hypothesis, it means the data fits the expected distribution well within the margin of error allowed by your significance level.

By following these steps, you can interpret whether the observed data aligns with your expected theoretical model or if there are significant deviations.

---------------------------------------------------------------------------------------------------------------------

## **Explain the assumptions and limitations of chi-square tests.**

Chi-square tests are widely used statistical tests for categorical data, but they come with specific assumptions and limitations that should be kept in mind. Here’s a breakdown:

### Assumptions:

1. **Independence**: The observations should be independent of one another. This means that the occurrence of one observation does not affect the occurrence of another. In practice, this often means that each subject or item should only be in one category and not duplicated.
2. **Sample Size**: Chi-square tests require a sufficiently large sample size to ensure reliable results. Specifically, the expected frequency in each cell of the contingency table should be at least 5. If some expected frequencies are less than 5, the results of the test might not be reliable.
3. **Categorical Data**: The data being analyzed should be categorical, meaning that it falls into distinct categories. Chi-square tests are not appropriate for continuous data unless it is first categorized into groups.
4. **Mutually Exclusive Categories**: Each observation should fall into one and only one category. The categories should be mutually exclusive.

### Limitations:

1. **Sensitivity to Small Sample Sizes**: When sample sizes are small, the chi-square test may not be accurate. Small samples can lead to misleading results because the approximation to the chi-square distribution becomes less reliable.
2. **Not Suitable for Small Expected Frequencies**: If the expected frequency in any cell of the contingency table is less than 5, the chi-square test may not be valid. In such cases, alternative tests like Fisher's exact test might be more appropriate.
3. **Data Aggregation**: Chi-square tests require that data be aggregated into categorical bins. This means that some information might be lost if the data are inherently continuous and not naturally suited to categorization.
4. **Cannot Measure Strength of Association**: While chi-square tests can determine if there is an association between variables, they do not provide information about the strength or direction of the association.
5. **Assumption of Random Sampling**: The test assumes that the sample is randomly drawn from the population. If the sampling method is biased, the results may not be generalizable.

Understanding these assumptions and limitations helps in choosing the right statistical test and in interpreting the results appropriately.

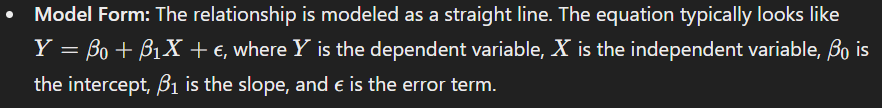
---------------------------------------------------------------------------------------------------------------------

## **What is the difference between simple linear regression and multiple regression?**

Simple linear regression and multiple regression are both statistical methods used to understand relationships between variables, but they differ in terms of complexity and the number of predictors involved.

**Simple Linear Regression:**

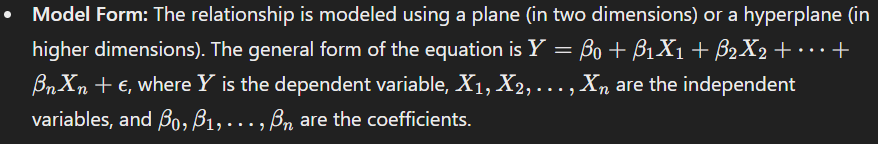
* **Number of Predictors:** It involves just one predictor (independent variable) and one outcome (dependent variable).

****

* **Usage:** It's used when you want to explore the relationship between two variables and predict the value of the dependent variable based on the independent variable.

**Multiple Regression:**

* **Number of Predictors:** It involves two or more predictors. This model can accommodate multiple independent variables and one dependent variable.

****

* **Usage:** It's used when you want to understand the relationship between the dependent variable and multiple independent variables, and also to see how each predictor contributes to the outcome, potentially controlling for the influence of other predictors.

In summary, simple linear regression examines the relationship between two variables, while multiple regression assesses the impact of several variables on a single outcome, providing a more nuanced view of how multiple factors interact to influence the dependent variable.

---------------------------------------------------------------------------------------------------------------------

## **How do you assess the goodness-of-fit of a regression model?**

Assessing the goodness-of-fit of a regression model involves evaluating how well the model's predictions match the actual data. Here are some common methods and metrics used for this purpose:

1. **R-squared (R^2)**: R^2 is a statistical measure that tells you how well the independent variables in a regression model explain the variation in the dependent variable. Values range from 0 to 1, where a higher value indicates a better fit. However, a high R^2 doesn’t necessarily mean the model is good, especially if it’s overfitting.

Imagine you're trying to predict someone's weight based on their height. If your prediction is perfect, R^2 would be 1. If your prediction is completely off, R^2 would be 0.

1. **Adjusted R-squared**: Adjusted R^2 is a modified version of R^2 that takes into account the number of independent variables in the model. It adjusts for the fact that adding more variables can artificially inflate the R^2 value. It’s useful for comparing models with different numbers of predictors, as it penalizes the addition of less significant predictors.

Suppose you add more and more details (like hair color, shoe size, etc.) to predict someone's weight. Regular R^2 might keep increasing just because you're adding more details. Adjusted R^2, on the other hand, will go down if those details don't actually help you make better predictions.

1. **Mean Absolute Error (MAE)**: This is the average of the absolute differences between the predicted and actual values. It gives a clear measure of prediction accuracy but doesn’t penalize large errors more heavily.
2. **Mean Squared Error (MSE)**: This is the average of the squared differences between the predicted and actual values. It gives more weight to larger errors compared to MAE, which can be useful if large errors are particularly undesirable.
3. **Root Mean Squared Error (RMSE)**: This is the square root of MSE. It has the same units as the dependent variable, making it easier to interpret than MSE.
4. **Residual Plots**: Plotting the residuals (differences between observed and predicted values) can help diagnose problems with the model. Residuals should be randomly scattered around zero if the model fits well. Patterns in residuals might indicate issues like non-linearity or heteroscedasticity.
5. **F-test**: In the context of linear regression, the F-test can be used to determine if the model is significantly better than a model with no predictors (i.e., a model that only uses the mean of the dependent variable).
6. **Cross-validation**: Techniques like k-fold cross-validation involve partitioning the data into subsets, training the model on some subsets, and testing it on others. This helps assess how well the model generalizes to unseen data.

Each method provides different insights, and often, a combination of these metrics is used to assess the overall performance of a regression model comprehensively.

## **Explain multicollinearity and how you would detect and handle it in a regression model.**

Multicollinearity occurs in regression analysis when two or more predictor variables (independent variables) are highly correlated with each other. This makes it difficult to determine the individual effect of each predictor on the dependent variable and can lead to unreliable estimates of regression coefficients.

### **Detection of Multicollinearity**

1. **Correlation Matrix**: Check the correlation matrix of the predictor variables. High correlations (usually above 0.8 or below -0.8) between predictors may indicate multicollinearity.
2. **Variance Inflation Factor (VIF)**: Calculate the VIF for each predictor. A VIF value greater than 10 (some use 5 as a threshold) suggests significant multicollinearity. VIF measures how much the variance of the estimated regression coefficients is inflated due to multicollinearity.
3. **Condition Index**: Compute the condition index, which is derived from the eigenvalues of the predictor variable matrix. A condition index greater than 30 may indicate multicollinearity.
4. **Eigenvalues**: Examine the eigenvalues of the correlation matrix. Small eigenvalues close to zero suggest multicollinearity.

### **Handling Multicollinearity**

1. **Remove Highly Correlated Predictors**: If two predictors are highly correlated, consider removing one of them from the model.
2. **Combine Predictors**: Use principal component analysis (PCA) or factor analysis to combine correlated predictors into a single predictor.
3. **Regularization Techniques**: Apply regularization methods such as Ridge Regression (L2 regularization) or Lasso Regression (L1 regularization). These techniques add a penalty to the regression model to reduce the impact of multicollinearity.
4. **Centering the Predictors**: Subtract the mean of each predictor to center them around zero. This can sometimes reduce multicollinearity, especially when dealing with interaction terms.
5. **Increase Sample Size**: Sometimes increasing the sample size can help mitigate the effects of multicollinearity.
6. **Check Data Quality**: Ensure there are no data entry errors or outliers that might be inflating correlations among predictors.

Addressing multicollinearity involves a balance between improving model stability and maintaining interpretability. The choice of method depends on the specific context and goals of your analysis.

---------------------------------------------------------------------------------------------------------------------

## **What is the difference between correlation and causation?**

The difference between correlation and causation is fundamental in understanding relationships between variables:

1. **Correlation**:
   * **Definition**: Correlation refers to a statistical relationship between two variables. It indicates that as one variable changes, the other tends to change in a specific pattern. This relationship can be positive (both variables increase or decrease together) or negative (one variable increases as the other decreases).
   * **Example**: There might be a correlation between ice cream sales and drowning incidents. As ice cream sales increase, drowning incidents might also increase. This doesn't mean that buying ice cream causes drowning; rather, both might be influenced by a third variable, such as hot weather.
2. **Causation**:
   * **Definition**: Causation implies that one variable directly affects another. In other words, a change in one variable directly leads to a change in the other variable.
   * **Example**: If you increase the amount of water a plant receives, it will grow faster. Here, the amount of water is causing the plant to grow faster.

To determine causation, researchers need to conduct experiments or studies that control for other variables and establish a direct cause-and-effect relationship. Correlation alone does not imply causation, and mistakenly assuming causation from correlation can lead to incorrect conclusions.

---------------------------------------------------------------------------------------------------------------------

## **How do you interpret the Pearson correlation coefficient?**

The Pearson correlation coefficient, often denoted as r, measures the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where:

* **1** indicates a perfect positive linear relationship: as one variable increases, the other variable increases proportionally.
* **-1** indicates a perfect negative linear relationship: as one variable increases, the other variable decreases proportionally.
* **0** indicates no linear relationship: changes in one variable do not predict changes in the other variable.

In general:

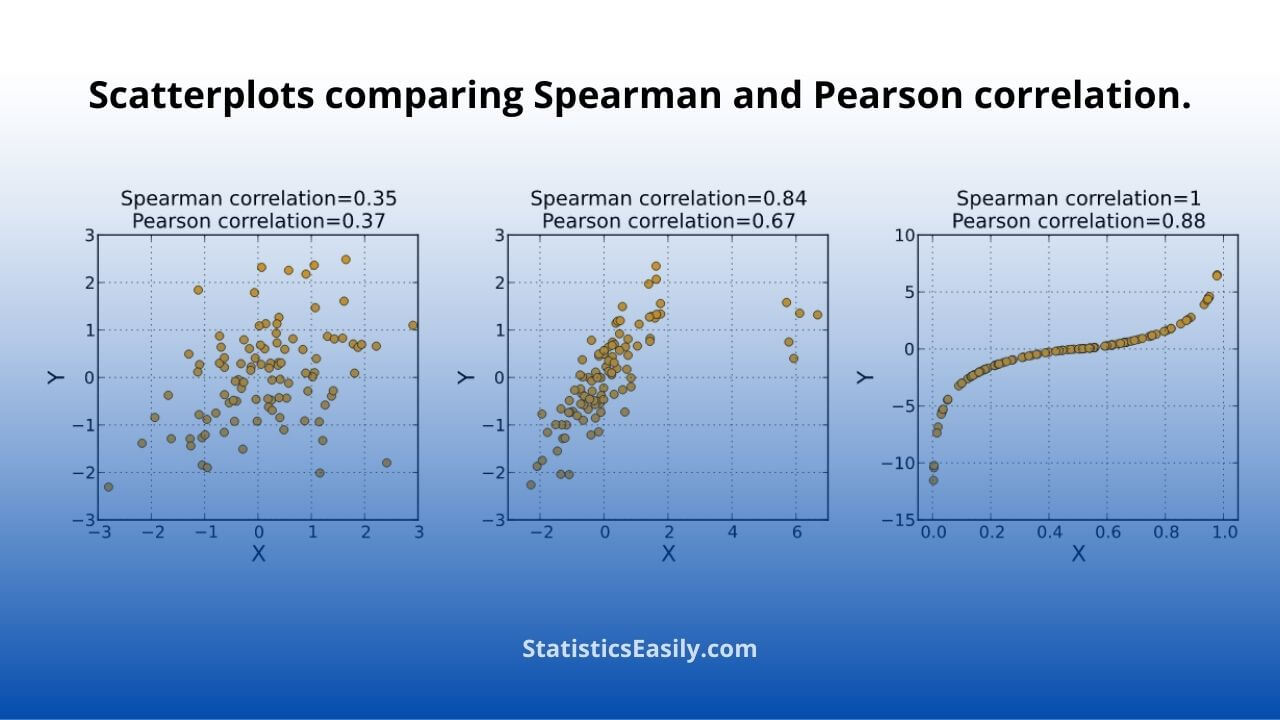
* **0.1 to 0.3** (or **-0.1 to -0.3**): Small or weak correlation
* **0.3 to 0.5** (or **-0.3 to -0.5**): Moderate correlation
* **0.5 to 1.0** (or **-0.5 to -1.0**): Strong correlation

It's important to remember that correlation does not imply causation. A strong correlation between two variables doesn’t necessarily mean that one causes the other.

---------------------------------------------------------------------------------------------------------------------

## **When would you use Spearman rank correlation instead of Pearson correlation?**

Spearman rank correlation is used instead of Pearson correlation when the **data doesn't meet the assumptions required for Pearson correlation**. Here are some specific situations where Spearman might be preferred:



1. **Non-linearity**: When the **relationship between variables is not linear**, Spearman's correlation can be more appropriate. **Pearson correlation measures linear relationships**, while Spearman assesses monotonic relationships (i.e., if one variable increases, the other variable tends to increase or decrease, but not necessarily at a constant rate).
2. **Ordinal Data**: Spearman's rank correlation is suitable for **ordinal data**, where the **variables represent ranks** rather than actual quantities. For instance, if you're ranking preferences or survey responses, Spearman's method is appropriate.
3. **Non-Normality**: When the data is **not normally distributed** or when there are **outliers**, Spearman's correlation is more robust. Spearman ranks the data and calculates correlation based on these ranks, which mitigates the impact of extreme values.
4. **Non-Interval Scales**: If the data is on a non-interval scale, where the distances between values are not uniform, Spearman's correlation is better suited because it relies on ranks rather than actual values.

In summary, use Spearman rank correlation when you're dealing with ordinal data, non-linear relationships, non-normally distributed data, or outliers that could skew the results of Pearson correlation.

---------------------------------------------------------------------------------------------------------------------

## **What are some common methods for forecasting time series data?**

Forecasting time series data involves predicting future values based on historical data. Here are some common methods:

1. **Naive Methods**:
   * **Naive Forecast**: Uses the last observed value as the forecast for all future periods.
   * **Seasonal Naive Forecast**: Uses the value from the same season in the previous cycle (e.g., last month or last year) as the forecast.
2. **Moving Averages**:
   * **Simple Moving Average (SMA)**: Averages the values over a specified number of past periods.
   * **Weighted Moving Average (WMA)**: Assigns different weights to past observations, giving more importance to recent data.
3. **Exponential Smoothing**:
   * **Simple Exponential Smoothing**: Applies a smoothing factor to the most recent observation and the previous forecast.
   * **Holt’s Linear Trend Model**: Extends simple exponential smoothing to account for trends in the data.
   * **Holt-Winters Seasonal Model**: Adds seasonal components to Holt’s model, suitable for data with seasonality.
4. **Autoregressive Integrated Moving Average (ARIMA)**:
   * **ARIMA Model**: Combines autoregressive (AR) terms, differencing (I) to make the series stationary, and moving average (MA) terms. Suitable for univariate time series data.
   * **Seasonal ARIMA (SARIMA)**: Extends ARIMA to handle seasonality in the data.
5. **Autoregressive Integrated Moving Average with Exogenous Regressors (ARIMAX)**:
   * Similar to ARIMA but includes external variables (exogenous regressors) to improve forecasts.
6. **Vector Autoregression (VAR)**:
   * Used for multivariate time series where multiple time series are interdependent.
7. **State Space Models**:
   * **Kalman Filter**: A recursive algorithm that estimates the state of a dynamic system from a series of incomplete and noisy measurements.
   * **Dynamic Linear Models (DLM)**: Uses state space representations for time series forecasting.
8. **Machine Learning Methods**:
   * **Regression Trees**: Model time series data using tree-based algorithms.
   * **Support Vector Machines (SVM)**: Can be adapted for forecasting with time series data.
   * **Neural Networks**: Includes methods like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRU) which are effective for capturing complex patterns in time series data.
9. **Prophet**:
   * Developed by Facebook, Prophet is designed to handle daily observations and can accommodate holidays and other special events.
10. **Bayesian Methods**:
    * **Bayesian Structural Time Series (BSTS)**: Provides probabilistic forecasts and incorporates different structural components such as trend and seasonality.

The choice of method depends on the characteristics of your data, such as seasonality, trend, and the presence of external variables.

---------------------------------------------------------------------------------------------------------------------

## **Explain the components of a time series (trend, seasonality, residuals).**

In time series analysis, understanding the different components can help you interpret and model the data more effectively. Here’s a breakdown of the main components:

1. **Trend**:
   * **Definition**: The trend represents the long-term movement or direction in the data. It's the underlying tendency for the data to increase, decrease, or stay constant over time.
   * **Example**: If you’re looking at monthly sales data for a company, a trend might show that sales are gradually increasing over several years.
2. **Seasonality**:
   * **Definition**: Seasonality refers to regular, repeating patterns or fluctuations in the data that occur at specific intervals, such as daily, weekly, monthly, or quarterly. These patterns are often influenced by seasonal factors or events.
   * **Example**: Retail sales often spike during holiday seasons, like December, showing a yearly seasonal pattern.
3. **Residuals** (or **Irregular Component**):
   * **Definition**: Residuals represent the **random noise or irregularities in the data** that can’t be attributed to the trend or seasonality. These are the unpredictable variations or errors that remain after removing the trend and seasonal components.
   * **Example**: After accounting for the trend and seasonal effects in monthly sales data, the residuals might include sudden, unexplained spikes or drops due to unusual events, like a local promotion or supply chain issues.

In practice, when analyzing a time series, you often decompose it into these components to better understand and forecast future values. Techniques like additive or multiplicative decomposition can help in this process.

---------------------------------------------------------------------------------------------------------------------

## **How would you handle missing data in a time series dataset?**

Handling missing data in a time series dataset can be a bit tricky because you want to maintain the temporal structure of the data. Here are several common strategies:

1. **Imputation**:
   * **Forward Fill**: Replace missing values with the last known value. This is useful when the missing data is expected to be similar to the previous observations.
   * **Backward Fill**: Replace missing values with the next known value. This can be useful if the missing data is likely to be similar to future observations.
   * **Linear Interpolation**: Estimate missing values by interpolating between the known values before and after the missing data. This is often useful for time series with a trend.
   * **Spline Interpolation**: A more sophisticated form of interpolation that fits a spline curve to the data points, which can be useful for more complex trends.
2. **Time Series Specific Methods**:
   * **Seasonal Decomposition**: Decompose the time series into seasonal, trend, and residual components, and impute the missing values within these components.
   * **Kalman Filter**: Use a Kalman filter or similar state-space model to estimate missing values based on the model's predictions.
3. **Statistical Methods**:
   * **Mean/Median Imputation**: Replace missing values with the mean or median of the observed data. This is a simple method but might not capture temporal dependencies well.
   * **Model-Based Imputation**: Use statistical models like ARIMA (AutoRegressive Integrated Moving Average) to predict and impute missing values based on the patterns in the data.
4. **Machine Learning Methods**:
   * **k-Nearest Neighbors (k-NN)**: Impute missing values based on similar observations in the dataset.
   * **Regression Models**: Predict missing values using regression models where the target variable is the missing value and the features are other time series data.
5. **Drop Missing Data**:
   * If the amount of missing data is very small, you might opt to simply remove those data points, especially if they don't significantly impact the analysis or modeling.
6. **Modeling Considerations**:
   * **Handling Missing Data in Models**: Some models can handle missing data inherently (e.g., certain machine learning algorithms). Check if your model has built-in mechanisms for dealing with missing values.

The choice of method depends on the nature of your data, the amount of missing data, and the context of your analysis. It’s often a good idea to try different methods and validate their impact on your model or analysis results.

---------------------------------------------------------------------------------------------------------------------

## **Describe your approach to exploratory data analysis (EDA).**

Exploratory Data Analysis (EDA) is a crucial step in understanding and preparing data before diving into more complex analyses or modeling. My approach to EDA generally involves the following steps:

1. **Understand the Context**:
   * **Objective**: Clarify the goals of the analysis. What questions are we trying to answer?
   * **Data Source**: Know where the data comes from and any relevant domain knowledge.
2. **Initial Data Assessment**:
   * **Data Loading**: Import the data and check the basic structure (rows, columns).
   * **Summary Statistics**: Generate descriptive statistics (mean, median, standard deviation) to get an overview of numerical features.
   * **Data Types**: Identify and confirm data types for each column (e.g., integer, float, categorical).
3. **Data Cleaning**:
   * **Missing Values**: Identify and handle missing data (e.g., imputation, removal).
   * **Duplicates**: Check for and address duplicate records.
   * **Outliers**: Detect and consider how to manage outliers.
4. **Univariate Analysis**:
   * **Distribution**: Plot histograms or density plots to understand the distribution of individual features.
   * **Categorical Variables**: Use bar charts to visualize the frequency of categories.
5. **Bivariate/Multivariate Analysis**:
   * **Correlation**: Compute and visualize correlations between numerical features using heatmaps or pair plots.
   * **Relationships**: Use scatter plots, box plots, or violin plots to explore relationships between variables.
6. **Data Visualization**:
   * **Visual Exploration**: Create various plots (e.g., scatter plots, line plots, bar charts) to visually inspect the data and identify patterns or anomalies.
   * **Feature Engineering**: Based on insights, create new features or transform existing ones.
7. **Data Transformation**:
   * **Normalization/Standardization**: Scale features if necessary for further analysis or modeling.
   * **Encoding Categorical Data**: Convert categorical variables into numerical format using techniques like one-hot encoding or label encoding.
8. **Preliminary Findings**:
   * **Insights**: Summarize key findings and patterns discovered during the analysis.
   * **Next Steps**: Identify areas requiring further investigation or more sophisticated modeling techniques.

Throughout the process, the goal is to gain a thorough understanding of the data’s structure, patterns, and potential issues to ensure that any subsequent analysis or modeling is built on a solid foundation.

---------------------------------------------------------------------------------------------------------------------

## **How do you handle outliers in a dataset?**

Handling outliers in a dataset depends on the context and the goals of your analysis. Here are some common approaches:

1. **Identify Outliers**:
   * **Statistical Methods**: Use techniques like Z-scores, IQR (Interquartile Range), or Tukey's fences to detect outliers.
   * **Visual Methods**: Create plots such as box plots, scatter plots, or histograms to visually inspect data for anomalies.
2. **Assess the Impact**:
   * Determine whether the outliers are errors, extreme but valid values, or indicative of a different distribution.
   * Consider the domain context to understand if the outliers have meaningful implications.
3. **Decide on a Strategy**:
   * **Remove Outliers**: If they are errors or irrelevant, you can exclude them from the analysis. This is often done when the outliers skew results significantly.
   * **Transform Data**: Apply transformations (e.g., logarithmic or square root) to reduce the influence of outliers.
   * **Cap or Winsorize**: Replace outliers with a specified percentile value (e.g., capping extreme values at the 5th and 95th percentiles).
   * **Imputation**: Replace outliers with estimated values based on other observations in the dataset.
   * **Separate Analysis**: Treat outliers separately to understand their impact on the data or to perform different analyses.
4. **Model Robustness**:
   * Use robust statistical methods or models that are less sensitive to outliers (e.g., median instead of mean, robust regression techniques).
5. **Validate**:
   * Check how your decisions affect the analysis and whether the results align with domain knowledge or practical considerations.

Handling outliers effectively requires balancing between preserving the integrity of the data and ensuring accurate, meaningful analysis results.

---------------------------------------------------------------------------------------------------------------------

## **Explain the steps you would take to validate the results of your analysis.**

Validating the results of an analysis is crucial to ensure accuracy and reliability. Here’s a step-by-step guide to validate your results:

1. **Verify Data Quality**:
   * **Check for Accuracy**: Ensure that the data used in the analysis is correct and accurately recorded.
   * **Assess Completeness**: Confirm that the data set is complete and doesn’t have missing or incomplete values.
   * **Identify and Handle Outliers**: Detect and address any outliers that might skew the results.
2. **Review Methodology**:
   * **Reevaluate Assumptions**: Check if the assumptions made during the analysis are valid and appropriate.
   * **Inspect the Analysis Process**: Ensure that the correct statistical methods and algorithms were used for the analysis.
3. **Reproduce Results**:
   * **Perform Replication**: Re-run the analysis using the same data and methodology to see if you obtain the same results.
   * **Use Alternative Methods**: Apply different methods or tools to analyze the same data and compare the results.
4. **Cross-Validation**:
   * **Split Data**: Divide the data into training and testing sets (if applicable) to validate the results on unseen data.
   * **Check Consistency**: Ensure that the results are consistent across different subsets of data.
5. **Compare with Benchmark**:
   * **Use Known Standards**: Compare the results with established benchmarks or known values to verify accuracy.
   * **Benchmark Against Previous Studies**: See if the results align with findings from similar analyses or studies.
6. **Conduct Sensitivity Analysis**:
   * **Test Variations**: Examine how sensitive the results are to changes in data or assumptions.
   * **Assess Robustness**: Determine if the results hold up under different scenarios or conditions.
7. **Seek Peer Review**:
   * **Get Feedback**: Have others review your analysis and results to identify any errors or oversights.
   * **Incorporate Expertise**: Consult with experts in the field to validate the findings and interpretations.
8. **Document and Communicate**:
   * **Record Methodology and Findings**: Document the steps, assumptions, and results thoroughly.
   * **Communicate Clearly**: Present the results in a clear and understandable manner, including any limitations or uncertainties.

By following these steps, you can ensure that your analysis is robust, accurate, and reliable.

---------------------------------------------------------------------------------------------------------------------

## **Give an example of how you have used statistical analysis to solve a real-world problem**

Sure! One real-world example of using statistical analysis is in the field of public health, specifically in tracking and managing the spread of infectious diseases. For instance, during the COVID-19 pandemic, statistical models were crucial for understanding and predicting the virus's spread.

Researchers used statistical analysis to:

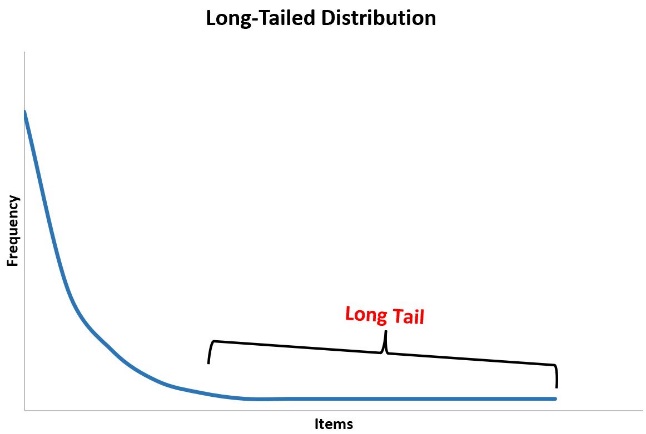
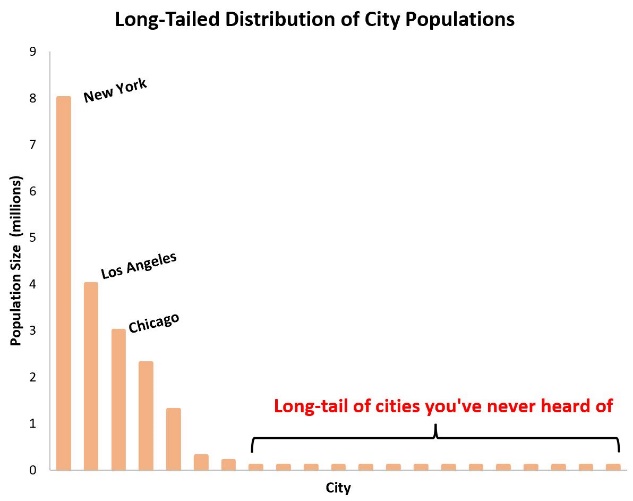
1. **Track Infection Rates**: By analyzing daily case numbers and testing data, they were able to estimate infection rates and identify trends in different regions.
2. **Predict Future Spread**: Using models like the SIR (Susceptible, Infected, Recovered) model or more complex variants, statisticians could forecast future infection rates under different scenarios, helping policymakers make informed decisions about interventions.
3. **Evaluate Interventions**: Statistical methods were employed to assess the effectiveness of public health measures like lockdowns, mask mandates, and vaccination campaigns. By comparing infection rates before and after these measures, analysts could gauge their impact.
4. **Resource Allocation**: Statistical analysis helped determine where to allocate resources such as ventilators and hospital beds based on projected needs, which was critical for managing healthcare capacity.

Overall, statistical analysis provided valuable insights that guided public health responses and helped mitigate the impact of the pandemic.

---------------------------------------------------------------------------------------------------------------------

## **Where are long-tailed distributions used?**

A long-tailed distribution is a type of distribution where the tail drops off gradually toward the end of the curve.



The Pareto principle and the **product sales distribution** are good examples to denote the use of long-tailed distributions. Also, it is widely used in [classification](https://intellipaat.com/blog/what-is-data-classification/) and [regression](https://intellipaat.com/blog/what-is-regression/) problems.

---------------------------------------------------------------------------------------------------------------------

## **What is observational and experimental data in Statistics?**

**Observational data** [correlates](https://intellipaat.com/blog/what-is-correlation-in-statistics/) to the data that is obtained from **observational studies**, where variables are observed to see if there is **any correlation between them.**

**Experimental data** is derived from experimental studies, where certain variables are held constant to see if any discrepancy is raised in the working.

---------------------------------------------------------------------------------------------------------------------

## **What is meant by mean imputation for missing data? Why is it bad?**

Mean imputation is a rarely used practice where null values in a dataset are replaced directly with the corresponding mean of the data.

It is considered a bad practice as it **completely removes the accountability for feature correlation**. This also means that the data will have **low variance and increased bias**, adding to the dip in the accuracy of the model, alongside narrower **confidence intervals**.

---------------------------------------------------------------------------------------------------------------------

## **What is an outlier? How can outliers be determined in a dataset?**

Outliers are data points that vary in a large way when compared to other observations in the dataset. Depending on the learning process, an outlier can worsen the accuracy of a model and decrease its efficiency sharply.

**Outliers are determined by using two methods:**

* Standard deviation/z-score
* Interquartile range (IQR)

---------------------------------------------------------------------------------------------------------------------

## **How is missing data handled in statistics?**

There are many ways to handle missing data in Statistics:

* Prediction of the missing values
* Assignment of individual (unique) values
* Deletion of rows, which have the missing data
* Mean imputation or median imputation
* Using [random forests](https://intellipaat.com/blog/what-is-random-forest-algorithm-in-python/), which support the missing values

---------------------------------------------------------------------------------------------------------------------

## **What is the meaning of selection bias?**

Selection bias happens when the people or data chosen for a study are not a true reflection of the whole group you want to learn about. This can lead to incorrect results because the study might only show what's true for the selected group, not everyone else.

**There are many types of selection bias as shown below:**

* Observer selection
* Attrition
* Protopathic bias
* Time intervals
* Sampling bias

---------------------------------------------------------------------------------------------------------------------

## **What is the probability of getting a sum of 5 or 8 when 2 dice are rolled once?**

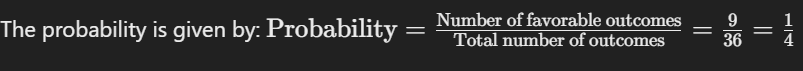
When 2 dice are rolled,

Total outcomes = 36 (i.e. 6\*6)

Possible outcomes of getting 5 = 4

Possible outcomes of getting a sum 8 = 5

Total = 9



Probability =9/36 = 1/4 = 0.25

---------------------------------------------------------------------------------------------------------------------

## **State the case where the median is a better measure when compared to the mean.**

In the case where there are a lot of outliers that can [positively or negatively skew](https://intellipaat.com/blog/kurtosis-and-skewness/) data, the median is preferred as it provides an accurate measure in this case of determination.

---------------------------------------------------------------------------------------------------------------------

## **Can you give an example of root cause analysis?**

Root cause analysis, as the name suggests, is a method used to solve problems by first identifying the root cause of the problem.

**Example**: If the higher crime rate in a city is directly associated with the higher sales in a red-colored shirt, it means that they are having a positive correlation. However, this does not mean that one causes the other.

Causation can always be tested using A/B testing or hypothesis testing.

---------------------------------------------------------------------------------------------------------------------

## **What type of data does not have a log-normal distribution or a Gaussian distribution?**

**Exponential distributions** do not have a log-normal distribution or a Gaussian distribution. In fact, any type of data that is categorical will not have these distributions as well.

Example: Duration of a phone call, time until the next earthquake, etc.

---------------------------------------------------------------------------------------------------------------------

## **What are quantitative data and qualitative data?**

* [Quantitative data](https://intellipaat.com/blog/what-is-quantitative-methods/) is also known as numeric data.
* Qualitative data is also known as categorical data.

---------------------------------------------------------------------------------------------------------------------

## **What are the types of sampling in Statistics?**

There are four main types of data sampling as shown below:

* **Simple random**: Pure random division
* **Cluster**: Population divided into clusters
* **Stratified**: Data divided into unique groups
* **Systematical**: Picks up every ‘n’ member in the data

---------------------------------------------------------------------------------------------------------------------

## **If a distribution is skewed to the right and has a median of 20, will the mean be greater than or less than 20?**

If the given distribution is a **right-skewed distribution**, then the mean should be greater than 20, while the mode remains to be less than 20.

---------------------------------------------------------------------------------------------------------------------

## **What is Bessel's correction?**

Bessel’s correction is a factor that is used to estimate a populations’ standard deviation from its sample. It causes the standard deviation to be less biased, thereby, providing more accurate results.

---------------------------------------------------------------------------------------------------------------------

## **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?**

First, correlation does not imply causation here. Correlation is only used to measure the relationship, which is linear between rest and productive work. If both vary rapidly, then it means that there is a high amount of correlation between them.

---------------------------------------------------------------------------------------------------------------------

## **What is the relationship between the confidence level and the significance level in statistics?**

**Confidence Level** refers to how certain you are that your sample data accurately reflects the true population parameter. For example, a 95% confidence level means you are 95% sure that the true value lies within your confidence interval.

**Significance Level** (denoted as alpha, α\alphaα) is the probability of rejecting the null hypothesis when it is actually true. It’s the threshold for deciding whether your findings are statistically significant. A common significance level is 0.05, meaning there’s a 5% chance of making a Type I error (false positive).

The **relationship** is that the confidence level is 1−1 -1− significance level. For example, a 95% confidence level corresponds to a 5% significance level.

---------------------------------------------------------------------------------------------------------------------

## **What types of variables are used for Pearson’s correlation coefficient?**

Variables to be used for the Pearson’s correlation coefficient must be either in a ratio or in an interval.

Note that there can exist a condition when one variable is a ratio, while the other is an interval score.

---------------------------------------------------------------------------------------------------------------------

## **In a scatter diagram, what is the line that is drawn above or below the regression line called?**

The line that is drawn above or below the regression line in a [scatter diagram](https://intellipaat.com/blog/types-of-data-visualization/) is called the residual or also the prediction error.

---------------------------------------------------------------------------------------------------------------------

## **What are the examples of symmetric distribution?**

Symmetric distribution means that the data on the left side of the median is the same as the one present on the right side of the median.

There are many examples of symmetric distribution, but the following three are the most widely used ones:

* Uniform distribution
* Binomial distribution
* Normal distribution

---------------------------------------------------------------------------------------------------------------------

## **Where is inferential statistics used?**

Inferential statistics is used for several purposes, such as research, in which we wish to draw conclusions about a population using some sample data. This is performed in a variety of fields, ranging from government operations to quality control and quality assurance teams in multinational corporations.

---------------------------------------------------------------------------------------------------------------------

## **What is the relationship between mean and median in a normal distribution?**

In a normal distribution, the mean is equal to the median. To know if the distribution of a dataset is normal, we can just check the dataset’s mean and median.

---------------------------------------------------------------------------------------------------------------------

## **What is the difference between the Ist quartile, the IInd quartile, and the IIIrd quartile?**

Quartiles are used to describe the distribution of data by splitting data into three equal portions, and the boundary or edge of these portions are called quartiles.

That is,

* The lower quartile (Q1) is the 25th [percentile](https://intellipaat.com/blog/percentile/).
* The middle quartile (Q2), also called the median, is the 50th percentile.
* The upper quartile (Q3) is the 75th percentile.

---------------------------------------------------------------------------------------------------------------------

## **How do the standard error and the margin of error relate?**

The standard error and the margin of error are quite closely related to each other. In fact, the margin of error is calculated using the standard error. As the standard error increases, the margin of error also increases.

---------------------------------------------------------------------------------------------------------------------

## **What is one sample t-test?**

This T-test is a statistical hypothesis test in which we check if the mean of the sample data is statistically or significantly different from the population’s mean.

---------------------------------------------------------------------------------------------------------------------

## **Given a left-skewed distribution that has a median of 60, what conclusions can we draw about the mean and the mode of the data?**

Given that it is a left-skewed distribution, the mean will be less than the median, i.e., less than 60, and the mode will be greater than 60.

---------------------------------------------------------------------------------------------------------------------

## **What are the types of biases that we encounter while sampling?**

Sampling biases are errors that occur when taking a small sample of data from a large population as the representation in statistical analysis. There are three types of biases:

* The selection bias
* The survivorship bias
* The undercoverage bias

---------------------------------------------------------------------------------------------------------------------

## **What are the scenarios where outliers are kept in the data?**

There are not many scenarios where outliers are kept in the data, but there are some important situations when they are kept. They are kept in the data for analysis if:

* Results are critical
* Outliers add meaning to the data
* The data is highly skewed

---------------------------------------------------------------------------------------------------------------------

## **What are some of the properties of a normal distribution?**

A normal distribution, regardless of its size, will have a bell-shaped curve that is symmetric along the axes.

Following are some of the important properties:

* **Unimodal**: It has only one mode.
* **Symmetrical**: Left and right halves of the curve are mirrored.
* **Central tendency**: The mean, median, and mode are at the midpoint.

---------------------------------------------------------------------------------------------------------------------

## **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)?**

**The probability of not seeing a supercar in 20 minutes is:**

|  |  |
| --- | --- |
| 1  2  3 | = 1 − P(Seeing one supercar)  = 1 − 0.3  = 0.7 |

**Probability of not seeing any supercar in the period of 60 minutes is:**

|  |  |
| --- | --- |
| 1 | = (0.7) ^ 3 = 0.343 |

**Hence, the probability of seeing at least one supercar in 60 minutes is:**

|  |  |
| --- | --- |
| 1  2 | = 1 − P(Not seeing any supercar)  = 1 − 0.343 = 0.657 |

---------------------------------------------------------------------------------------------------------------------

## **What are some of the low and high-bias Machine Learning algorithms?**

There are many low and high-bias Machine Learning algorithms, and the following are some of the widely used ones:

Low bias: [SVM](https://intellipaat.com/blog/tutorial/machine-learning-tutorial/svm-algorithm-in-python/), decision trees, KNN algorithm, etc.

High bias: Linear and logistic regression

---------------------------------------------------------------------------------------------------------------------

## **What is the use of Hash tables in statistics?**

[Hash tables](https://intellipaat.com/blog/what-is-hash-table/) are the data structures that are used to denote the representation of key-value pairs in a structured way. The hashing function is used by a hash table to compute an index that contains all of the details regarding the keys that are mapped to their associated values.

---------------------------------------------------------------------------------------------------------------------

## **What is the benefit of using box plots?**

Box plots allow us to provide a graphical representation of the 5-number summary and can also be used to compare groups of histograms.

---------------------------------------------------------------------------------------------------------------------

## **Does a symmetric distribution need to be unimodal?**

A symmetric distribution does not need to be unimodal (having only one mode or one value that occurs most frequently). It can be bi-modal (having two values that have the highest frequencies) or multi-modal (having multiple or more than two values that have the highest frequencies).

---------------------------------------------------------------------------------------------------------------------

## **What is a survivorship bias?**

The survivorship bias is the flaw of the sample selection that occurs when a dataset only considers the ‘surviving’ or existing observations and fails to consider those observations that have already ceased to exist.

------------------------------------------------------------------------------------------------------------------------------------------

**What is an undercoverage bias?**

The undercoverage bias is a bias that occurs when some members of the population are inadequately represented in the sample.

Outliers in statistics have a very negative impact as they skew the result of any statistical query. For example, if we want to calculate the mean of a dataset that contains outliers, then the mean calculated will be different from the actual mean (i.e., the mean we will get once we remove the outliers).

---------------------------------------------------------------------------------------------------------------------

## **What is the relationship between standard deviation and standard variance?**

Standard deviation is the square root of standard variance. Basically, standard deviation takes a look at how the data is spread out from the mean. On the other hand, standard variance is used to describe how much the data varies from the mean of the entire dataset.

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------------------